



US006161172A

United States Patent [19][11] **Patent Number:** **6,161,172****Narayan et al.**[45] **Date of Patent:** **Dec. 12, 2000**

[54] **METHOD FOR CONCURRENTLY DISPATCHING MICROCODE AND DIRECTLY-DECODED INSTRUCTIONS IN A MICROPROCESSOR**

[75] **Inventors:** Rammohan Narayan; Rupaka Mahalingalah; Paul K. Miller, all of Austin, Tex.

[73] **Assignee:** Advanced Micro Devices, Inc., Sunnyvale, Calif.

[21] **Appl. No.:** 09/198,120

[22] **Filed:** Nov. 23, 1998

Related U.S. Application Data

[63] Continuation of application No. 08/878,228, Jun. 18, 1997, Pat. No. 5,884,058, which is a continuation of application No. 08/685,656, Jul. 24, 1996, abandoned.

[51] **Int. Cl.⁷** G06F 9/30

[52] **U.S. Cl.** 712/204; 712/206; 712/214; 712/215

[58] **Field of Search** 712/204, 206, 712/214, 215

[56] **References Cited****U.S. PATENT DOCUMENTS**

4,044,338 8/1977 Wolf 365/49
4,453,212 6/1984 Gaither et al. 711/2

(List continued on next page.)

FOREIGN PATENT DOCUMENTS

0259095 3/1988 European Pat. Off. .
0381471 8/1990 European Pat. Off. .

(List continued on next page.)

OTHER PUBLICATIONS

Intel, "Chapter 2: Microprocessor Architecture Overview," pp. 2-1 through 2-4.
Michael Slater, "AMD's K5 Designed to Outrun Pentium," Microprocessor Report, vol. 8, No. 14, Oct. 24, 1994, 7 pages.

Sebastian Rupley and John Clyman, "P6: The Next Step?," PC Magazine, Sep. 12, 1995, 16 pages.

Ton R. Halfhill, "AMD K6 Takes on Intel P6," BYTE, Jan. 1996, 4 pages.

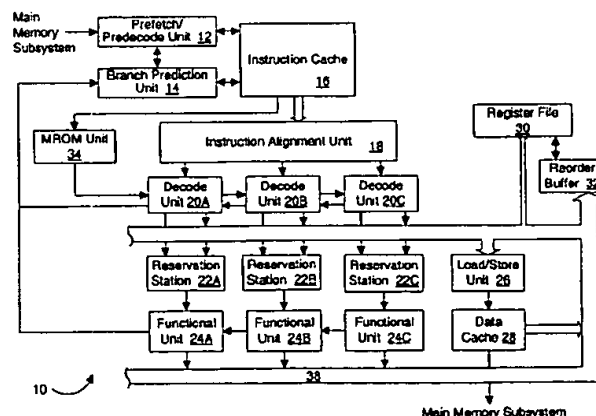
Primary Examiner—Ahmad F. Matar

Assistant Examiner—Saleh Najjar

Attorney, Agent, or Firm—Conley, Rose & Tayon, PC; Lawrence J. Merkel

[57] **ABSTRACT**

A method of instruction dispatch is provided in which a directly-decoded instruction and a microcode instruction are concurrently dispatched ("packed"). The instruction which is second in program order is retained until the succeeding clock cycle. During the succeeding clock cycle, a microcode unit determines if the microcode instruction and the directly-decoded instruction, when taken together, occupy less than or equal to the total number of issue positions available in the microprocessor. If the microcode unit determines that less than or equal to the total number of issue positions are occupied, then the packing is successful. If the microcode unit determines that greater than the total number of issue positions are occupied, then the packing is unsuccessful and the retained instruction is redispached. Additionally, instruction dispatch selection is performed in two phases. First, a number of instructions are selected as potentially dispatchable instructions. From the potentially dispatchable instructions, a set of actually dispatched instructions may be selected based upon the success or failure of instruction packing during the previous clock cycle and whether or not packing was performed. If instruction packing was not performed during the previous clock cycle or was performed unsuccessfully, then the instructions which are foremost in program order within the potentially dispatchable instructions are selected. However, if instruction packing was successfully performed in the previous clock cycle, then the retained instruction is not selected for dispatch.

19 Claims, 12 Drawing Sheets

U.S. PATENT DOCUMENTS

4,807,115	2/1989	Tomg	712/215	5,560,028	9/1996	Sachs et al.	712/23
4,839,797	6/1989	Katori et al.	712/210	5,566,298	10/1996	Boggs et al.	712/10
4,858,105	8/1989	Kuriyama et al.	712/235	5,600,806	2/1997	Brown et al.	712/204
5,115,500	5/1992	Larsen	712/209	5,619,666	4/1997	Coon et al.	712/208
5,167,026	11/1992	Murray et al.	712/210	5,625,787	4/1997	Mahin et al.	712/204
5,202,967	4/1993	Matsuzaki et al.	712/212	5,630,083	5/1997	Carbine et al.	712/212
5,226,126	7/1993	McFarland et al.	712/218	5,655,097	8/1997	Witt et al.	712/204
5,226,130	7/1993	Favor et al.	712/238	5,664,134	9/1997	Galhup et al.	712/235
5,233,696	8/1993	Suzuki	712/204	5,689,672	11/1997	Witt et al.	712/213
5,235,686	8/1993	Bosshart	712/247	5,696,955	12/1997	Goddard et al.	712/222
5,337,415	8/1994	DeLano et al.	712/213	5,748,978	5/1998	Narayan et al.	712/23
5,371,864	12/1994	Chuang	712/206	5,758,114	5/1998	Johanson et al.	712/204
5,394,558	2/1995	Arakawa et al.	712/211	5,781,789	7/1998	Narayan	712/23
5,394,559	2/1995	Arakawa	455/5.1	5,819,057	10/1998	Witt et al.	712/204
5,430,851	7/1995	Hirata et al.	712/212	5,822,559	10/1998	Narayan et al.	712/214
5,459,844	10/1995	Eickemeyer et al.	712/213	5,826,071	10/1998	Narayan	712/224
5,488,710	1/1996	Sato et al.	711/125	5,832,249	11/1998	Tran et al.	712/204
5,488,729	1/1996	Vegesna et al.	712/209				
5,500,942	3/1996	Eickemeyer et al.	712/210				
5,504,923	4/1996	Ando	712/23				
5,509,130	4/1996	Trauben et al.	712/215				
5,559,975	9/1996	Christie et al.	712/230				

FOREIGN PATENT DOCUMENTS

0459232	12/1991	European Pat. Off. .
2263985	8/1993	United Kingdom .
2263987	8/1993	United Kingdom .
2281422	3/1995	United Kingdom .

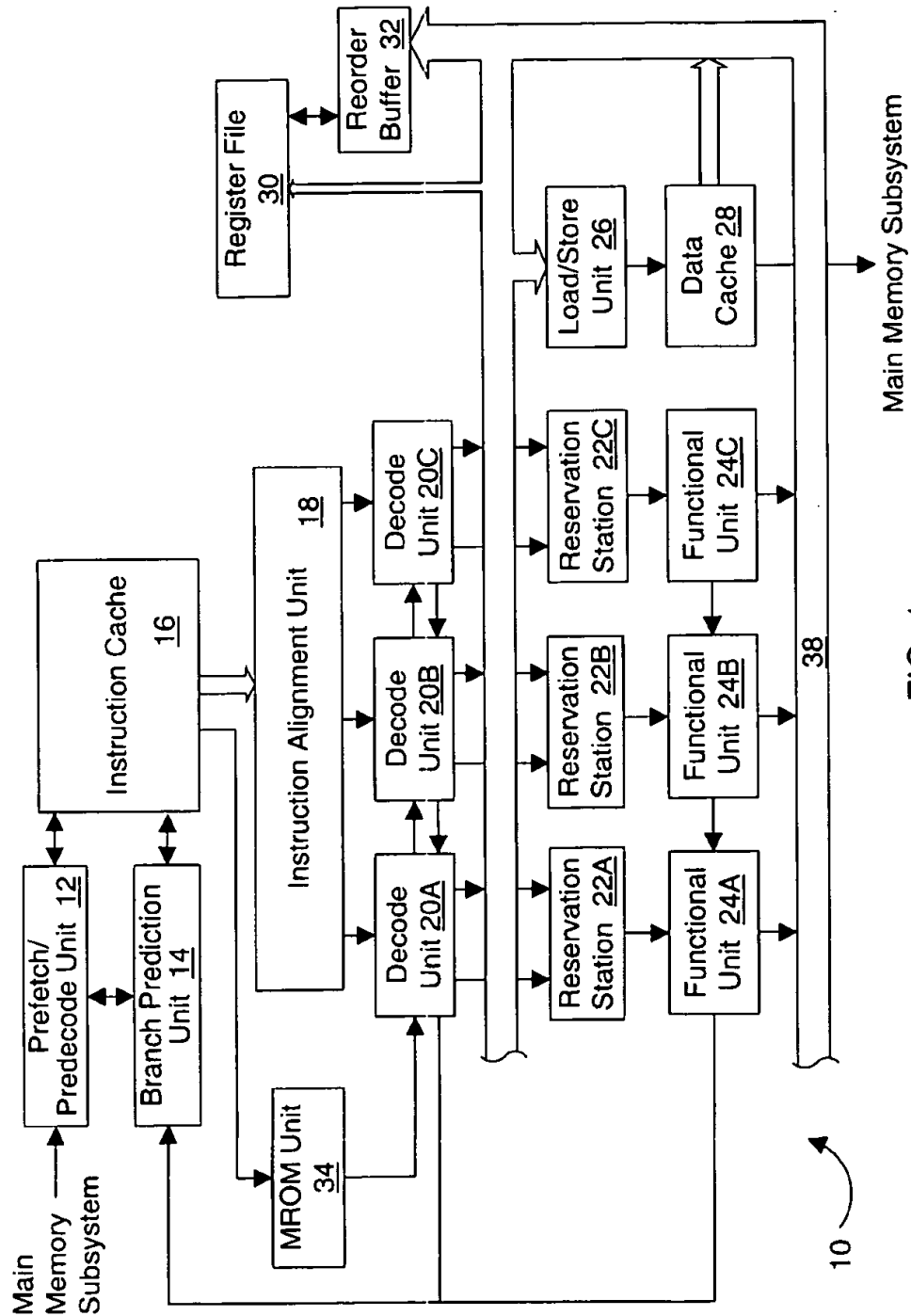


FIG. 1

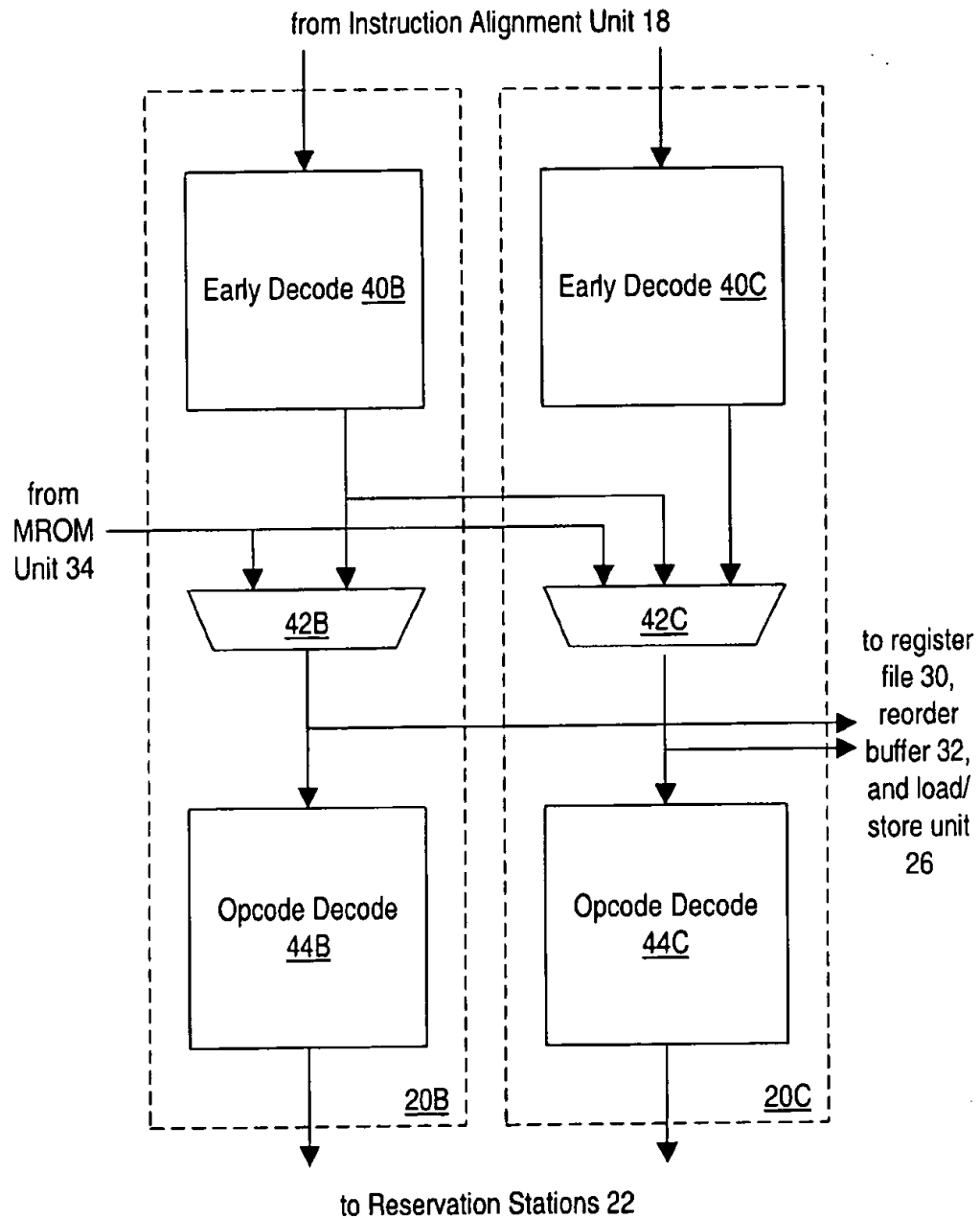


FIG. 2

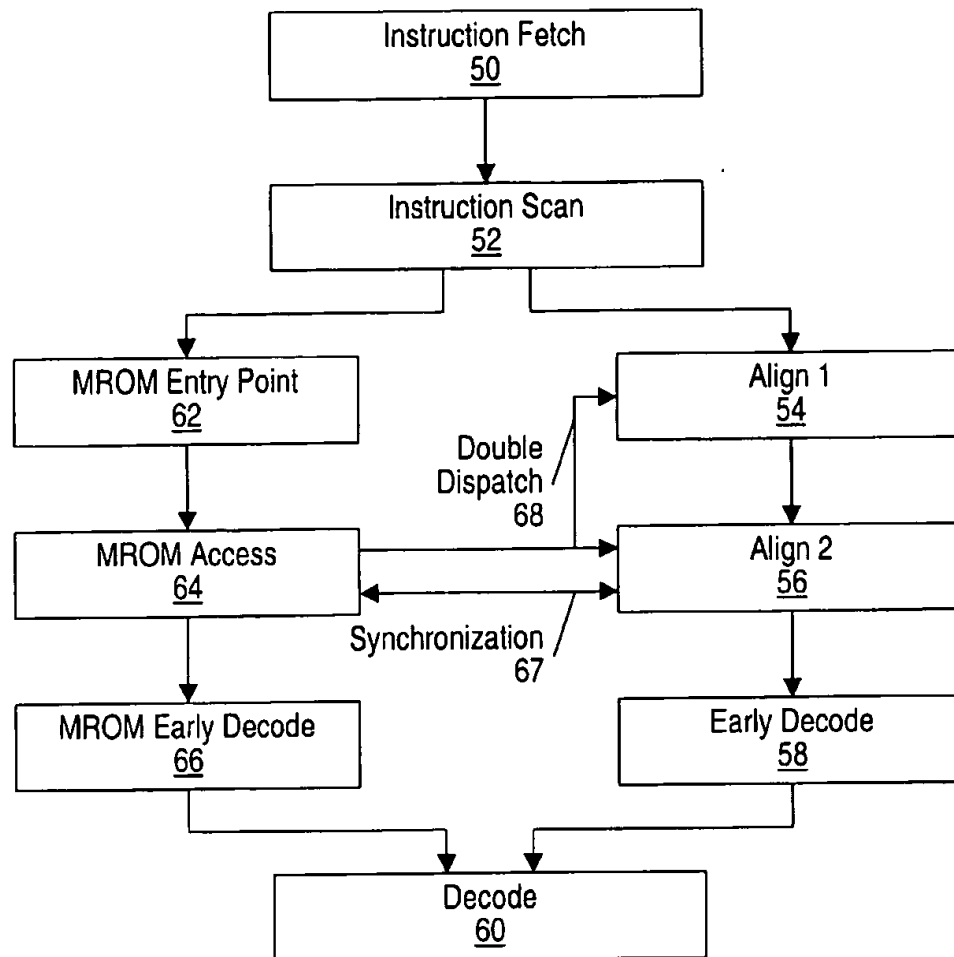


FIG. 3

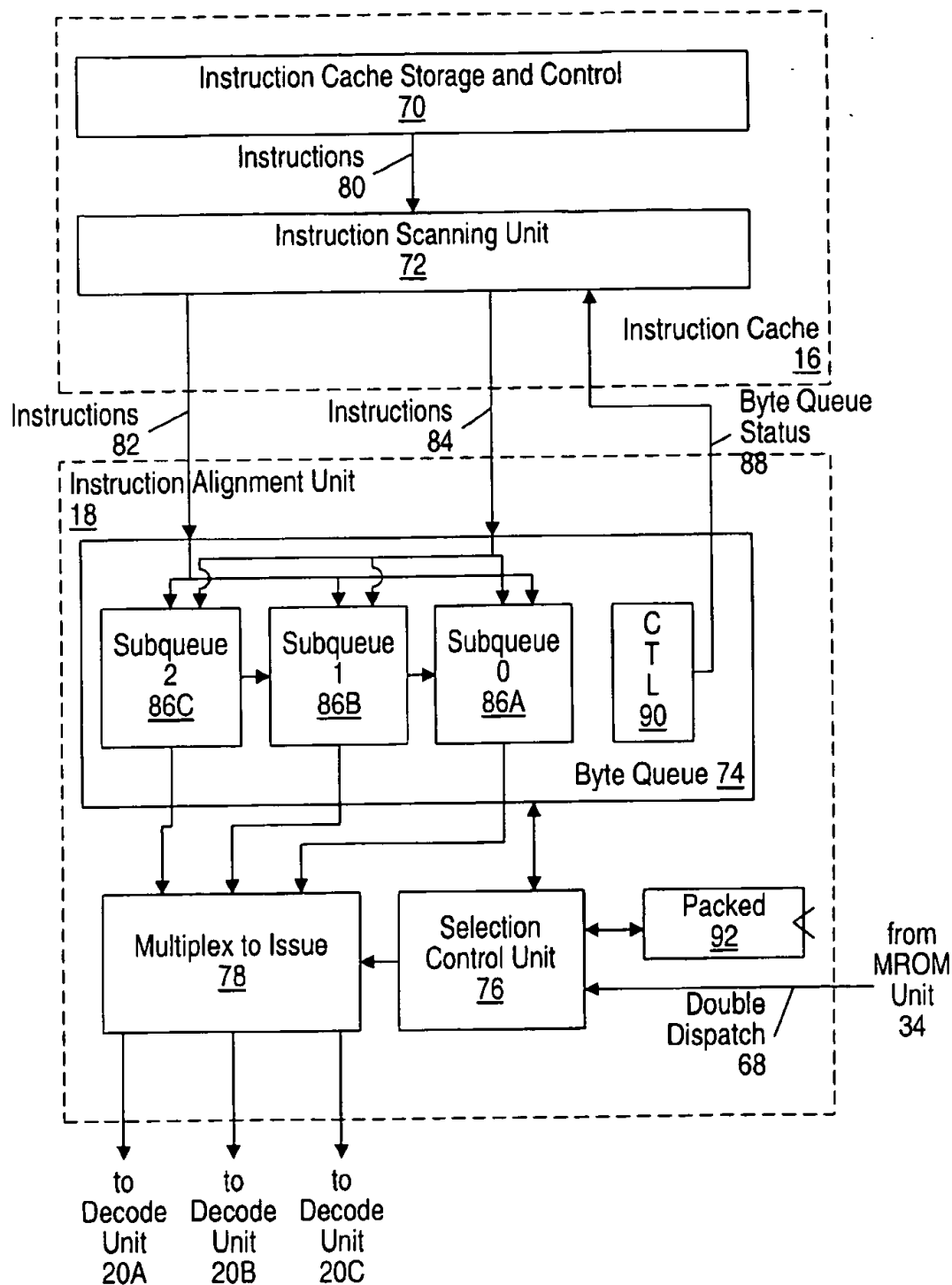


FIG. 4

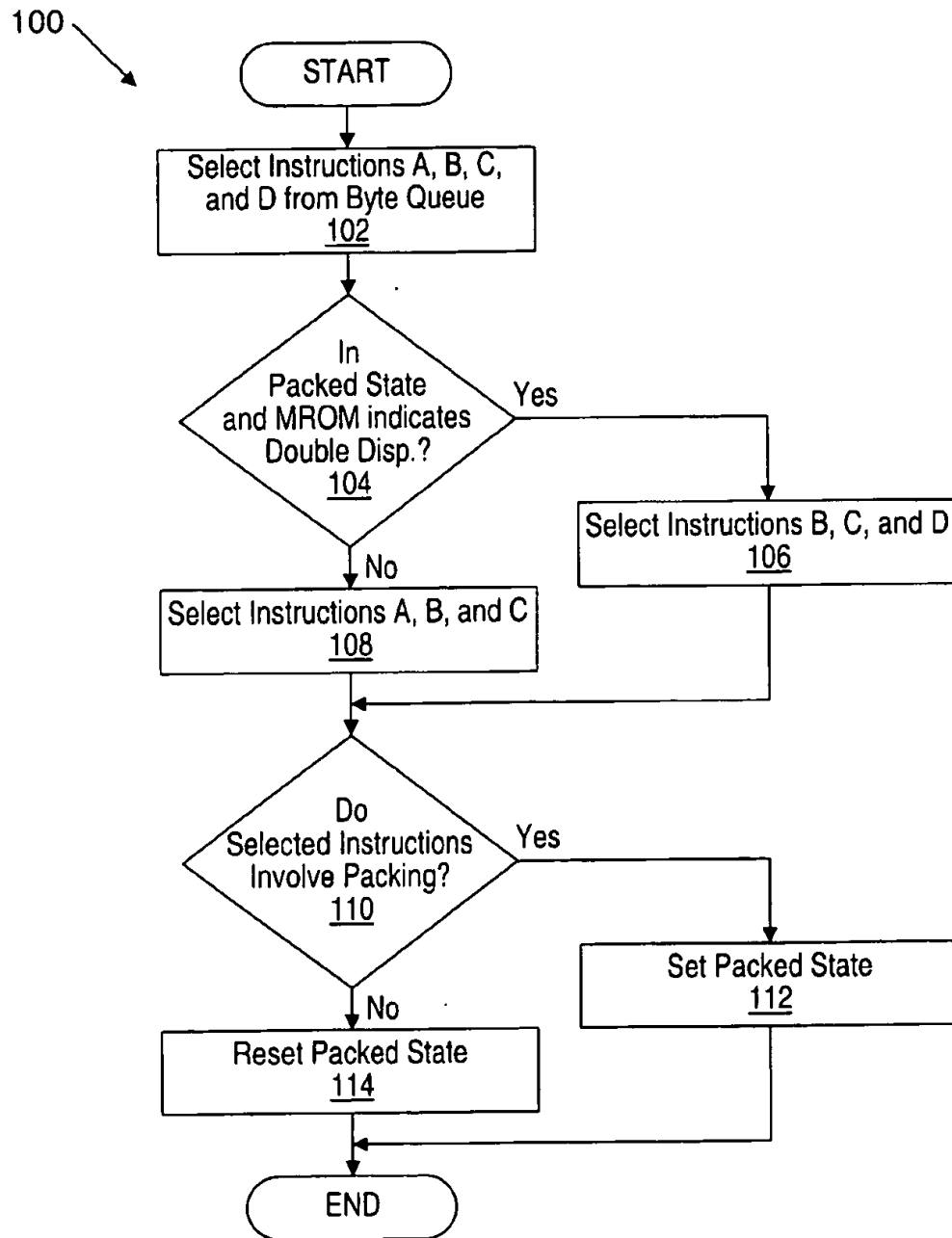


FIG. 5

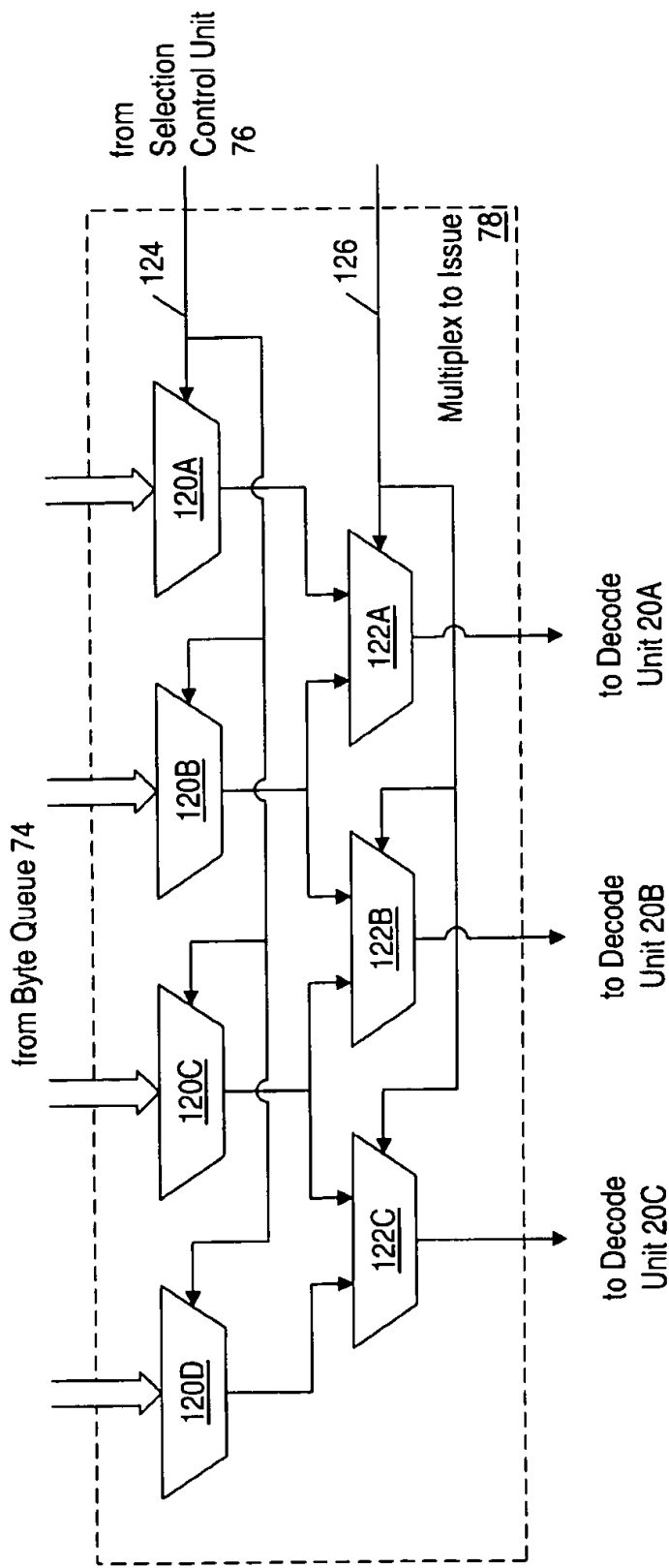


FIG. 6

Valid Combinations of
Instructions within a Subqueue

I2	I1	I0
X	X	X
X	X	A
X	B	A
C	B	A
O	X	X
O	X	A
O	B	A


130 

FIG. 7

Multiplexor	Positions Analyzed		
	Subqueue Zero 86A	Subqueue One 86B	Subqueue Two 86C
120A	I0, O	I0, O	I0
120B	I1, O	I0, I1, O	I0, I1
120C	I2, O	I0, I1, I2, O	I0, I1, I2
120D	—	I0, I1, I2, O	I0, I1, I2

FIG. 8

132

Final Issue Position Combinations

Issue Position 2	Issue Position 1	Issue Position 0
—	—	—
—	—	F
—	F	F
F	F	F
—	F	M
—	M	F
—	—	M

134

136

138

FIG. 9

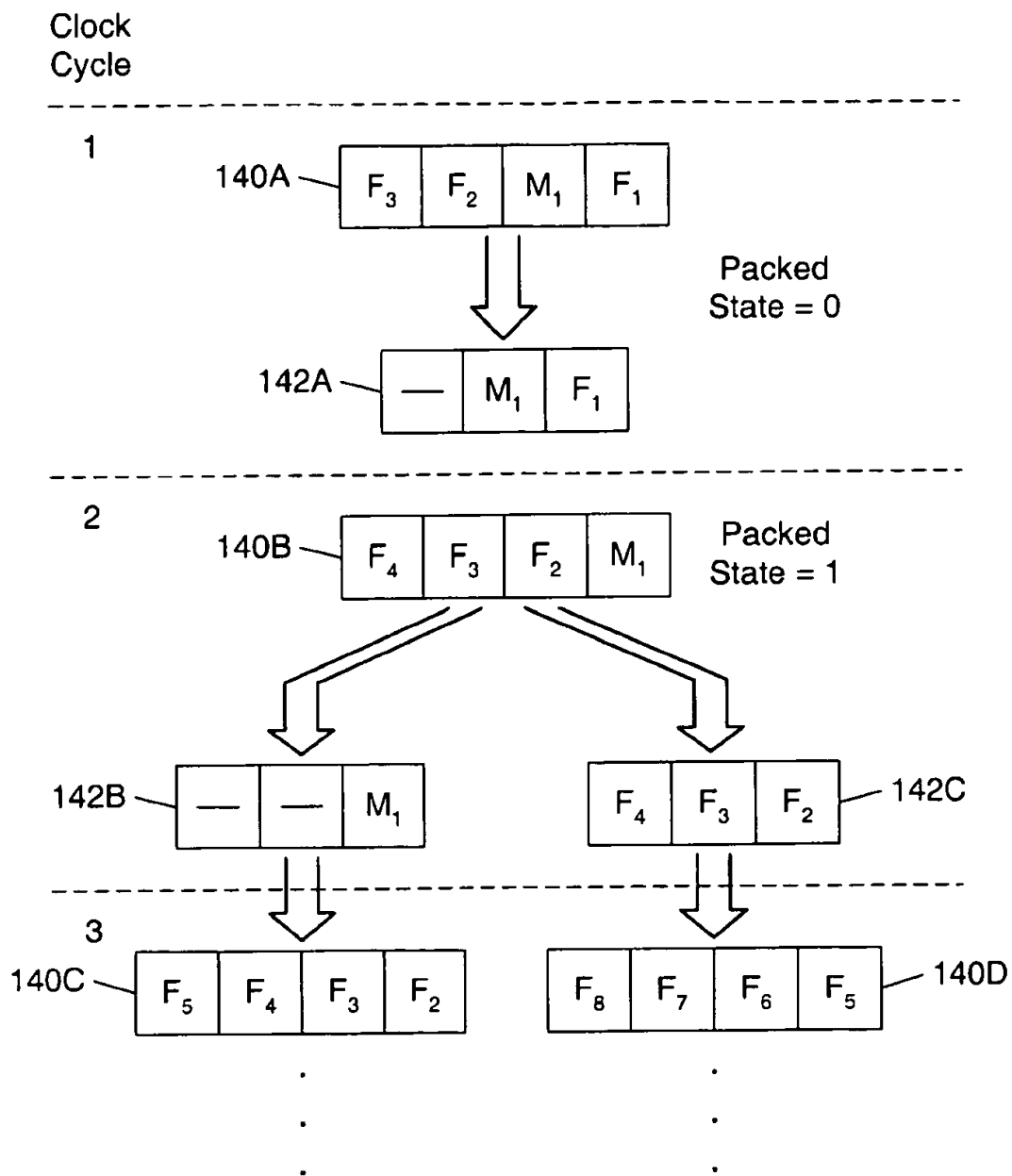


FIG. 10

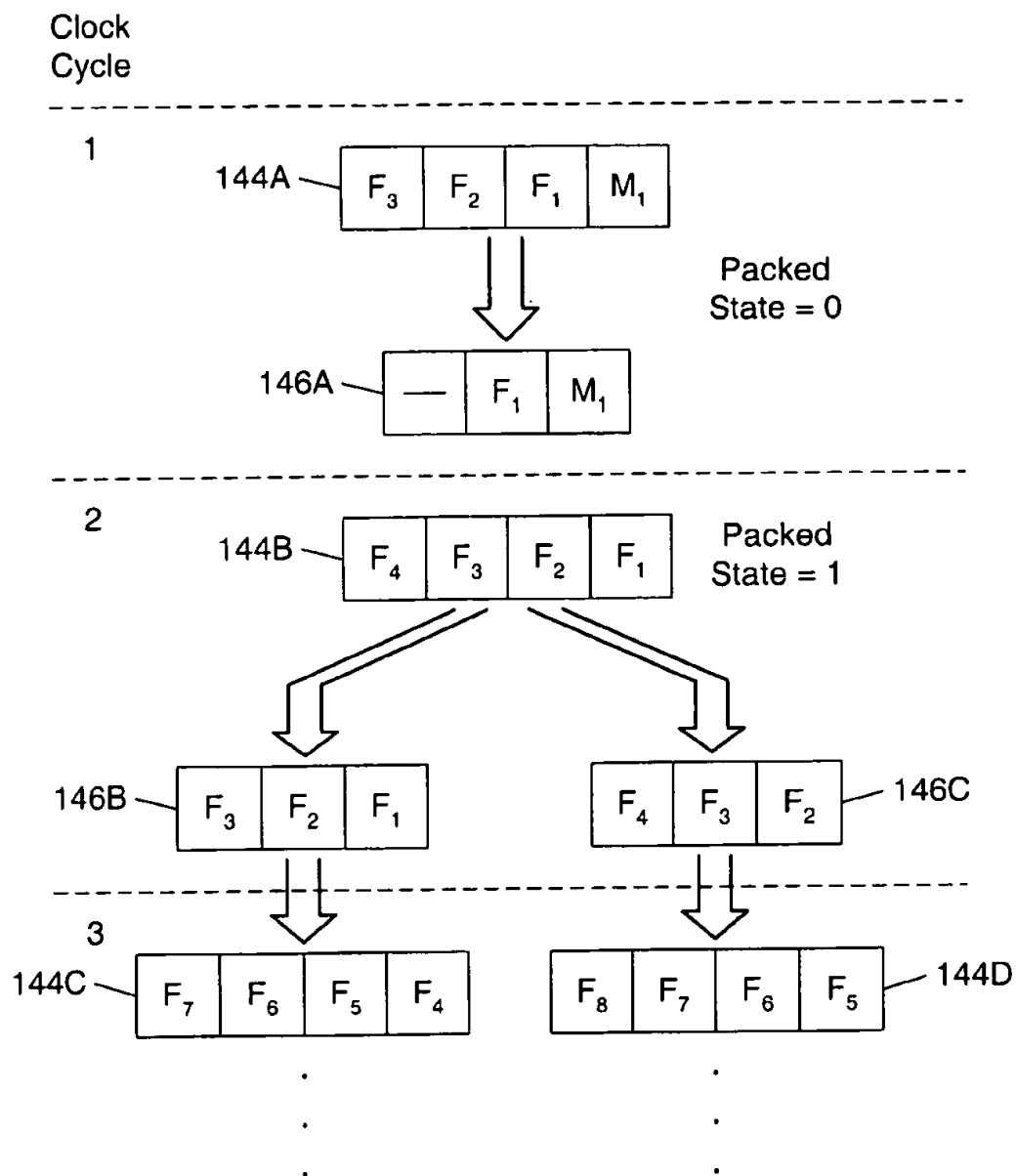


FIG. 11

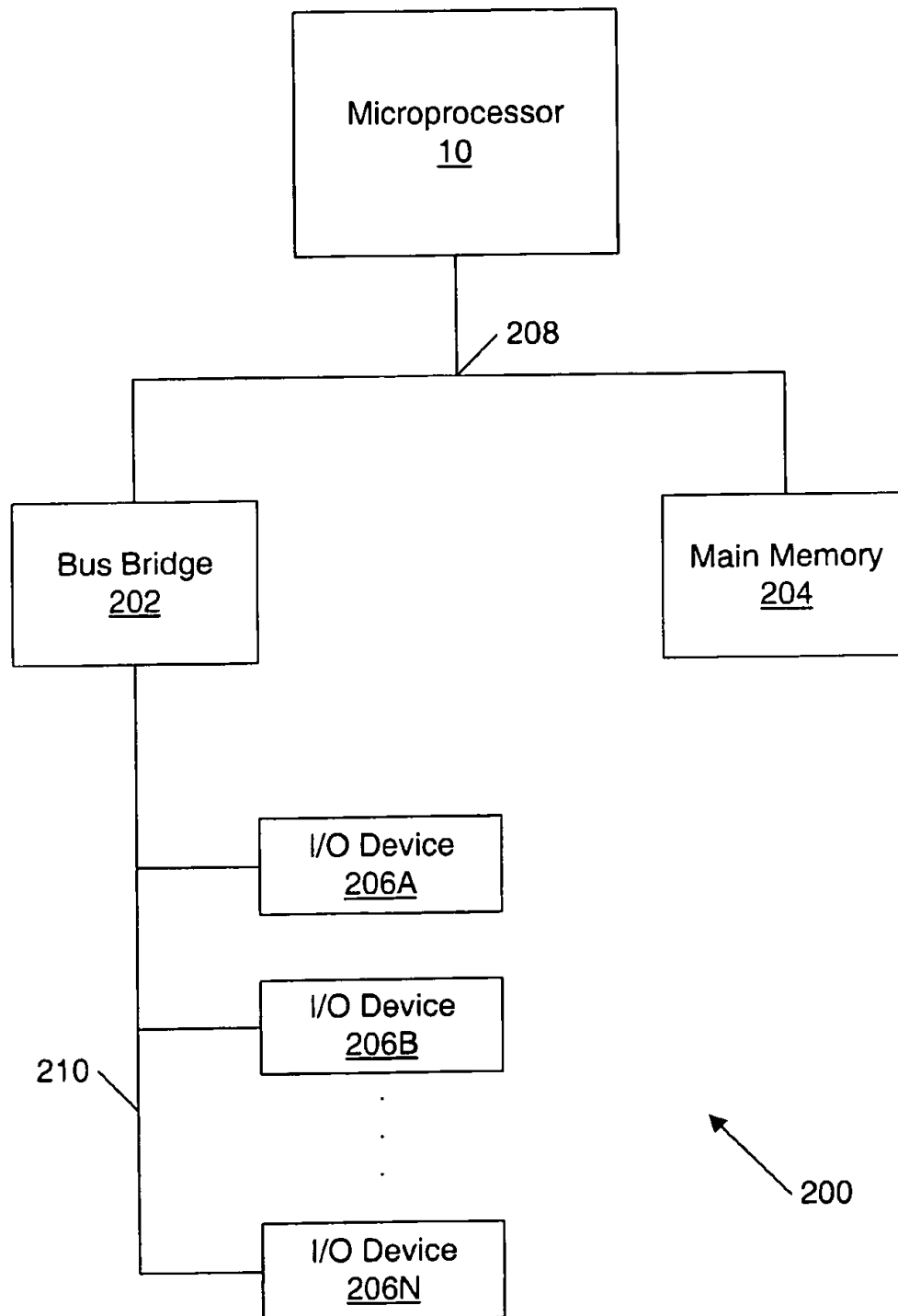


FIG. 12

METHOD FOR CONCURRENTLY DISPATCHING MICROCODE AND DIRECTLY-DECODED INSTRUCTIONS IN A MICROPROCESSOR

This application is a Continuation of U.S. Ser. No. 08/878,228, filed on Jun. 18, 1997 now U.S. Pat. No. 5,884,058 which is a File Wrapper Continuation of U.S. Ser. No. 08/685,656, filed Jul. 24, 1996, now Abn.

BACKGROUND OF THE INVENTION

1. Field of the Invention

This invention relates to the field of microprocessors and, more particularly, to instruction dispatch mechanisms within microprocessors.

2. Description of the Relevant Art

Superscalar microprocessors achieve high performance by executing multiple instructions per clock cycle and by choosing the shortest possible clock cycle consistent with the design. As used herein, the term "clock cycle" refers to an interval of time accorded to various stages of an instruction processing pipeline within the microprocessor. Storage devices (e.g. registers and arrays) capture their values according to the clock cycle. For example, a storage device may capture a value according to a rising or falling edge of a clock signal defining the clock cycle. The storage device then stores the value until the subsequent rising or falling edge of the clock signal, respectively. The term "instruction processing pipeline" is used herein to refer to the logic circuits employed to process instructions in a pipelined fashion. Although the pipeline may be divided into any number of stages at which portions of instruction processing are performed, instruction processing generally comprises fetching the instruction, decoding the instruction, executing the instruction, and storing the execution results in the destination identified by the instruction.

Microprocessor designers often design their products in accordance with the x86 microprocessor architecture in order to take advantage of its widespread acceptance in the computer industry. Because the x86 microprocessor architecture is pervasive, many computer programs are written in accordance with the architecture. X86 compatible microprocessors may execute these computer programs, thereby becoming more attractive to computer system designers who desire x86-capable computer systems. Such computer systems are often well received within the industry due to the wide range of available computer programs.

The x86 microprocessor architecture specifies a variable length instruction set (i.e. an instruction set in which various instructions employ differing numbers of bytes to specify that instruction). For example, the 80386 and later versions of x86 microprocessors employ between 1 and 15 bytes to specify a particular instruction. Instructions have an opcode, which may be 1-2 bytes, and additional bytes may be added to specify addressing modes, operands, and additional details regarding the instruction to be executed. Certain instructions within the x86 instruction set are quite complex, specifying multiple operations to be performed. For example, the PUSH instruction specifies that each of the x86 registers be pushed onto a stack defined by the value in the ESP register. The corresponding operations are a store operation for each register, and decrements of the ESP register between each store operation to generate the address for the next store operation.

Often, complex instructions are classified as microcode instructions. Microcode instructions are transmitted to a

microcode unit within the microprocessor, which decodes the complex microcode instruction and produces two or more simpler instructions for execution by the microprocessor. The simpler instructions corresponding to the microcode instruction are typically stored in a read-only memory (ROM) within the microcode unit. The microcode unit determines an address within the ROM at which the simpler instructions are stored, and transfers the instructions out of the ROM beginning at that address. Multiple clock cycles may be used to transfer the entire set of instructions corresponding to the microcode instruction. Each microcode instruction may correspond to a particular number of simpler instructions dissimilar from the number of simpler instructions corresponding to other microcode instructions. Additionally, the number of simpler instructions corresponding to a particular microcode instruction may vary according to the addressing mode of the instruction, the operand values, and/or the options included with the instruction. The microcode unit issues the simpler instructions into the instruction processing pipeline of the microprocessor. The simpler instructions are thereafter executed in a similar fashion to other instructions. It is noted that the simpler instructions may be instructions defined within the instruction set, or may be custom instructions defined for the particular microprocessor.

Conversely, less complex instructions are decoded by hardware decode units within the microprocessor, without intervention by the microcode unit. The term "directly-decoded instruction" will be used herein to refer to instructions which are decoded and executed by the microprocessor without the aid of a microcode unit. As opposed to microcode instructions which are reduced to simpler instructions which may be handled by the microprocessor, directly-decoded instructions are decoded and executed via hardware decode and functional units included within the microprocessor.

Unfortunately, having microcode instructions which translate to an arbitrary number of simpler instructions creates numerous problems for dispatching multiple instructions per clock cycle. Because the number of translated instructions is not known at the time of transmitting the microcode instruction to the microcode unit, instructions are typically not concurrently dispatched with the microcode instruction. Instead, the microcode instruction is typically dispatched alone, and subsequent dispatch is typically stalled until the simpler instructions corresponding to the microcode instructions have been dispatched. For cases in which the microcode instruction corresponds to a number of instructions less than the maximum number of instructions which may be dispatched during a clock cycle, dispatch bandwidth (i.e. the number of concurrently dispatched instructions) is wasted. Performance of the microprocessor may thereby be deleteriously reduced from the level achievable when dispatch bandwidth is fully utilized.

SUMMARY OF THE INVENTION

The problems outlined above are in large part solved by a method of instruction dispatch as described herein. According to the method a directly-decoded instruction and a microcode instruction are concurrently dispatched ("packed"). The instruction which is second in program order is retained until the succeeding clock cycle. During the succeeding clock cycle, a microcode unit determines if the microcode instruction and the directly-decoded instruction, when taken together, occupy less than or equal to the total number of issue positions available in the microprocessor. If the microcode unit determines that less than or equal to the

total number of issue positions are occupied, then the packing is successful. If the microcode unit determines that greater than the total number of issue positions are occupied, then the packing is unsuccessful and the retained instruction is redispached. Advantageously, instruction dispatch bandwidth is increased when packing is successful. Performance of a microprocessor employing the instruction dispatch method may be beneficially increased due to the increased utilization of processor resources as compared to microprocessors which issue microcode instructions without concurrent dispatch of other instructions. In other words, a larger average number of instructions executed per clock cycle may be achieved for instruction code which includes microcode instructions which do not occupy each of the available issue positions.

Additionally, instruction dispatch selection is performed in two phases. First, a number of instructions are selected as potentially dispatchable instructions. The number selected may be larger than the total number of available issue positions within the microprocessor. From the potentially dispatchable instructions, a set of actually dispatched instructions may be selected based upon the success or failure of instruction packing during the previous clock cycle. If instruction packing was not performed during the previous clock cycle, then the instructions which are foremost in program order within the potentially dispatchable instructions are selected. Similarly, if instruction packing was unsuccessfully performed during the previous clock cycle, the instructions which are foremost in program order within the potentially dispatchable instructions are selected. In this case, the first instruction in program order is the retained instruction. However, if instruction packing was successfully performed in the previous clock cycle, then the retained instruction is not selected for dispatch. Advantageously, instructions may be concurrently dispatched with the redispach of the retained instruction when instruction packing is unsuccessful. Instruction dispatch bandwidth is not sacrificed when redispach of a previously dispatched instruction becomes necessary. A microprocessor designed in accordance with the method may thereby achieve a maximal average number of instructions dispatched per clock cycle.

Broadly speaking, the present invention contemplates a method for dispatching instructions in a microprocessor having a plurality of issue positions comprising multiple steps. A first instruction and a second instruction are dispatched during a first clock cycle. The first instruction precedes the second instruction in program order, and one of the first instruction and the second instruction is a microcode instruction. The other one of the first instruction and the second instruction is a directly-decoded instruction. The second instruction and a third instruction are selected for dispatch during a second clock cycle subsequent to the first clock cycle. If the microcode instruction and the directly-decoded instruction, when dispatched together, occupy a first number of issue positions greater than a total number of the plurality of issue positions, the second instruction is dispatched. Alternatively, the third instruction is dispatched during the second clock cycle if the microcode instruction and the directly-decoded instruction, when dispatched together, occupy a second number of issue positions less than or equal to the total number of the plurality of issue positions.

The present invention further contemplates a method of dispatching instructions to a plurality of issue positions in a microprocessor comprising multiple steps. A state bit is set upon dispatch of a first instruction and a second instruction,

wherein one of the first instruction and the second instruction is a microcode instruction and the other is a directly-decoded instruction. The second instruction is redispached upon receipt of an indication that the microcode instruction occupies a first number of the plurality of issue positions which, when added to a second number of issue positions occupied by the directly-decoded instruction, exceeds a total number of the plurality of issue positions. The redispach occurs if the state bit is set.

The present invention still further contemplates a method for concurrently dispatching a microcode instruction and a directly-decoded instruction to a plurality of issue positions comprising several steps. A first plurality of instructions is dispatched during a first clock cycle, and includes the microcode instruction and the directly-decoded instruction. A second plurality of instructions is selected for dispatch during a second clock cycle subsequent to the first clock cycle. The second plurality of instructions is greater in number than a total number of the plurality of issue positions, and includes at least one of the first plurality of instructions. From the second plurality of instructions, at least one instruction is dispatched during the second clock cycle. The instruction is one of the first plurality of instructions if the microcode instruction is determined to occupy a first number of issue positions which, when added to a second number of issue positions occupied by a remainder of the first plurality of instructions, is greater than the total number of the plurality of issue positions. The instruction is not one of the first plurality of instructions if the first number added to the second number is less than or equal to the total number of issue positions.

BRIEF DESCRIPTION OF THE DRAWINGS

Other objects and advantages of the invention will become apparent upon reading the following detailed description and upon reference to the accompanying drawings in which:

FIG. 1 is a block diagram of one embodiment of a superscalar microprocessor.

FIG. 2 is a block diagram of one embodiment of a pair of decode units shown in FIG. 1.

FIG. 3 is a diagram depicting a portion of an instruction processing pipeline employed by one embodiment of the microprocessor shown in FIG. 1.

FIG. 4 is a block diagram of one embodiment of an instruction cache and an instruction alignment unit shown in FIG. 1.

FIG. 5 is a flowchart illustrating operation of the instruction alignment unit shown in FIG. 4 according to one embodiment of the instruction alignment unit.

FIG. 6 is a logic diagram of one embodiment of a multiplex to issue unit shown in FIG. 4.

FIG. 7 is a table depicting combinations of instructions which may be stored in a byte queue depicted in FIG. 4, according to one embodiment of the byte queue.

FIG. 8 is a table depicting instructions analyzed by a selection control unit shown in FIG. 4, according to one embodiment of the selection control unit.

FIG. 9 is a table of issue position combinations which are selected by the selection control unit shown in FIG. 4, according to one embodiment of the selection control unit.

FIG. 10 is an example of instruction selection.

FIG. 11 is a second example of instruction selection.

FIG. 12 is a block diagram of a computer system including the microprocessor shown in FIG. 1.

While the invention is susceptible to various modifications and alternative forms, specific embodiments thereof are shown by way of example in the drawings and will herein be described in detail. It should be understood, however, that the drawings and detailed description thereto are not intended to limit the invention to the particular form disclosed, but on the contrary, the intention is to cover all modifications, equivalents and alternatives falling within the spirit and scope of the present invention as defined by the appended claims.

DETAILED DESCRIPTION OF THE INVENTION

Turning now to FIG. 1, a block diagram of one embodiment of a microprocessor 10 is shown. Microprocessor 10 includes a prefetch/predecode unit 12, a branch prediction unit 14, an instruction cache 16, an instruction alignment unit 18, a plurality of decode units 20A-20C, a plurality of reservation stations 22A-22C, a plurality of functional units 24A-24C, a load/store unit 26, a data cache 28, a register file 30, a reorder buffer 32, and an MROM unit 34. Blocks referred to herein with a reference number followed by a letter will be collectively referred to by the reference number alone. For example, decode units 20A-20C will be collectively referred to as decode units 20.

Prefetch/predecode unit 12 is coupled to receive instructions from a main memory subsystem (not shown), and is further coupled to instruction cache 16 and branch prediction unit 14. Similarly, branch prediction unit 14 is coupled to instruction cache 16. Still further, branch prediction unit 14 is coupled to decode units 20 and functional units 24. Instruction cache 16 is further coupled to MROM unit 34 and instruction alignment unit 18. Instruction alignment unit 18 is in turn coupled to decode units 20. Each decode unit 20A-20C is coupled to load/store unit 26 and to respective reservation stations 22A-22C. Reservation stations 22A-22C are further coupled to respective functional units 24A-24C. Additionally, decode units 20 and reservation stations 22 are coupled to register file 30 and reorder buffer 32. Functional units 24 are coupled to load/store unit 26, register file 30, and reorder buffer 32 as well. Data cache 28 is coupled to load/store unit 26 and to the main memory subsystem. Finally, MROM unit 34 is coupled to decode units 20.

Generally speaking, microprocessor 10 categorizes microcode instructions as either double dispatch or arbitrary dispatch. Arbitrary dispatch microcode instructions may be dispatched to any number of issue positions, and are therefore dispatched without other instructions. In one particular embodiment, two subclasses of arbitrary dispatch are included: triple dispatch and more than triple dispatch. Conversely, double dispatch instructions occupy a pair of issue positions (i.e. double dispatch microcode instructions are parsed into a pair of simpler instructions). In the embodiment of FIG. 1, microprocessor 10 includes three issue positions. A double dispatch instruction does not occupy one of the issue positions. Therefore, a directly-decoded instruction may be dispatched concurrently with the double dispatch instructions. The directly-decoded instruction may be immediately prior to or immediately following the double dispatch instruction in program order. Advantageously, dispatch bandwidth is not wasted for cases in which double dispatch microcode instructions are encountered. Instead, the remaining issue position is filled with a directly-decoded instruction.

MROM unit 34 may detect double dispatch microcode instructions subsequent to the clock cycle in which instruc-

tions are selected for dispatch. Instructions are selected for dispatch based upon predecode information, which identifies microcode instructions as opposed to directly-decoded instructions but does not identify double dispatch versus arbitrary dispatch microcode instructions. The instruction dispatch selection logic within microprocessor 10 assumes that an MROM instruction selected for dispatch is a double-dispatch instruction, and therefore selects a directly-decoded instruction for concurrent dispatch. If MROM unit 34 detects that the microcode instruction is not a double dispatch instruction, then either the microcode instruction or the directly-decoded instruction is redispached during the following clock cycle (whichever one is second in program order). If the microcode instruction is redispached, it is redispached alone. If the directly-decoded instruction is redispached, it may be redispached along with subsequent instructions. Advantageously, even though microcode instructions are not classified until after dispatch, concurrent dispatch of microcode and directly-decoded instructions may be accomplished. Still further, subsequent dispatch bandwidth may not be wasted since redispached directly-decoded instructions may be dispatched concurrently with subsequent instructions.

Microprocessor 10 is configured to align instructions from instruction cache 16 to decode units 20 using instruction alignment unit 18. Instructions are fetched as an aligned plurality of bytes from a cache line within instruction cache 16. Instructions of interest may be stored beginning at any arbitrary byte within the fetched bytes. For example, a branch instruction may be executed having a target address which lies within a cache line. The instructions of interest therefore begin at the byte identified by the target address of the branch instruction. From the instruction bytes fetched, instruction alignment unit 18 identifies the instructions to be executed. Instruction alignment unit 18 conveys the instructions, in predicted program order, to decode units 20 for decode and execution.

Instruction alignment unit 18 includes a byte queue configured to store instruction bytes. An instruction scanning unit within instruction cache 16 separates the instructions fetched into instruction blocks. Each instruction block comprises a predefined number of instruction bytes. The instruction scanning unit identifies up to a predefined maximum number of instructions within the instruction block. Instruction identification information for each of the identified instructions is conveyed to instruction alignment unit 18 and is stored in the byte queue. The instruction identification information includes an indication of the validity of the instruction, as well as indications of the start and end of the instruction within the predefined number of instruction bytes. In one embodiment, the predefined number of instruction bytes comprises eight instruction bytes stored in contiguous main memory storage locations. The eight instruction bytes are aligned to an eight byte boundary (i.e. the least significant three bits of the address of the first of the contiguous bytes are zero). If more than the maximum number of instructions are contained within a particular predefined number of instruction bytes, the instruction bytes are scanned again during a subsequent clock cycle. The same instruction bytes are conveyed as another instruction block, with the additional instructions within the instruction bytes identified by the accompanying instruction identification information. Therefore, an instruction block may be defined as up to a predefined maximum number of instructions contained within a predefined number of instruction bytes.

The byte queue stores each instruction block and corresponding instruction identification information within a sub-

queue defined therein. The subqueues include a position for each possible valid instruction within the instruction block. The positions store instruction identification information and are maintained such that the instruction identification information for the first valid instruction within the subqueue is stored in a first position within the subqueue, instruction identification information regarding the second valid instruction (in program order) is stored in a second position within the subqueue, etc. When instructions within the subqueue are dispatched, instruction identification information corresponding to subsequent instructions are shifted within the positions of the subqueue such that the first of the remaining instructions is stored in the first position. Advantageously, instruction alignment unit 18 may only consider the instruction information stored in the first position of each subqueue to detect the instruction to be dispatched to decode unit 20A. Similarly, only the second position of the first subqueue (the subqueue storing instructions prior to the instructions stored in the other subqueues in program order) may be considered for dispatch of instructions to decode unit 20B. By managing the subqueues in this manner, logic for selecting and aligning instructions may be simplified. Fewer cascaded levels of logic may be employed for performing the selection and alignment process, allowing for high frequency implementation of microprocessor 10.

Because instructions are variable length, an instruction may begin within a particular instruction block but end in another instruction block. Instructions beginning within a particular instruction block and ending in another instruction block are referred to as "overflow instructions". The subqueue storing the instruction block within which an overflow instruction begins uses the last position to store the overflow instruction's identification information. Unlike the other positions, the instruction identification information of the last position is not shifted from the last position when an overflow instruction is stored therein. Advantageously, instruction alignment unit 18 need only search the last position of a particular subqueue to identify an instruction overflowing from one subqueue to another.

As used herein, the term queue refers to a storage device for storing a plurality of data items. The data items are stored with an ordered relationship between them. For example, the data items of the byte queue are instructions. The ordered relationship between the instructions is the program order of the instructions. Data items are removed from the queue according to the ordered relationship in a first-in-first-out (FIFO) fashion. Additionally, the term shifting is used to refer to movement of data items within the queue. When a data item is shifted from a first storage location to a second storage location, the data item is copied from the first storage location to the second storage location and invalidated in the second storage location. The invalidation may occur by shifting yet another data item into the second storage location, or by resetting a valid indication in the second storage location.

Instruction cache 16 is a high speed cache memory provided to store instructions. Instructions are fetched from instruction cache 16 and dispatched to decode units 20. In one embodiment, instruction cache 16 is configured to store up to 32 kilobytes of instructions in an 8 way set associative structure having 32 byte lines (a byte comprises 8 binary bits). Instruction cache 16 may additionally employ a way prediction scheme in order to speed access times to the instruction cache. Instead of accessing tags identifying each line of instructions and comparing the tags to the fetch address to select a way, instruction cache 16 predicts the way

that is accessed. In this manner, the way is selected prior to accessing the instruction storage. The access time of instruction cache 16 may be similar to a direct-mapped cache. A tag comparison is performed and, if the way prediction is incorrect, the correct instructions are fetched and the incorrect instructions are discarded. It is noted that instruction cache 16 may be implemented as a fully associative, set associative, or direct mapped configuration.

Instructions are fetched from main memory and stored into instruction cache 16 by prefetch/predecode unit 12. Instructions may be prefetched prior to instruction cache 16 recording a miss for the instructions in accordance with a prefetch scheme. A variety of prefetch schemes may be employed by prefetch/predecode unit 12. As prefetch/predecode unit 12 transfers instructions from main memory to instruction cache 16, prefetch/predecode unit 12 generates three predecode bits for each byte of the instructions: a start bit, an end bit, and a functional bit. The predecode bits form tags indicative of the boundaries of each instruction. The predecode tags may also convey additional information such as whether a given instruction can be decoded directly by decode units 20 or whether the instruction is executed by invoking a microcode procedure controlled by MROM unit 34, as will be described in greater detail below. Still further, prefetch/predecode unit 12 may be configured to detect branch instructions and to store branch prediction information corresponding to the branch instructions into branch prediction unit 14.

One encoding of the predecode tags for an embodiment of microprocessor 10 employing the x86 instruction set will next be described. If a given byte is the first byte of an instruction, the start bit for that byte is set. If the byte is the last byte of an instruction, the end bit for that byte is set. For this embodiment of microprocessor 10, instructions which may be directly decoded by decode units 20 are referred to as "fast path" instructions. Fast path instructions may be an example of directly-decoded instructions for this embodiment. The remaining x86 instructions are referred to as MROM instructions, according to one embodiment. For this embodiment, MROM instructions are an example of microcode instructions.

For fast path instructions, the functional bit is set for each prefix byte included in the instruction, and cleared for other bytes. Alternatively, for MROM instructions, the functional bit is cleared for each prefix byte and set for other bytes. The type of instruction may be determined by examining the functional bit corresponding to the end byte. If that functional bit is clear, the instruction is a fast path instruction. Conversely, if that functional bit is set, the instruction is an MROM instruction. The opcode of an instruction may thereby be located within an instruction which may be directly decoded by decode units 20 as the byte associated with the first clear functional bit in the instruction. For example, a fast path instruction including two prefix bytes, a Mod R/M byte, and an SIB byte would have start, end, and functional bits as follows:

Start bits	10000
End bits	00001
Functional bits	11000

MROM instructions are instructions which are determined to be too complex for decode by decode units 20. MROM instructions are executed by invoking MROM unit 34. More specifically, when an MROM instruction is

encountered, MROM unit 34 parses and issues the instruction into a subset of defined fast path instructions to effectuate the desired operation. MROM unit 34 dispatches the subset of fast path instructions to decode units 20. A listing of exemplary x86 instructions categorized as fast path instructions will be provided further below.

Microprocessor 10 employs branch prediction in order to speculatively fetch instructions subsequent to conditional branch instructions. Branch prediction unit 14 is included to perform branch prediction operations. In one embodiment, up to two branch target addresses are stored with respect to each cache line in instruction cache 16. Prefetch/predecode unit 12 determines initial branch targets when a particular line is predecoded. Subsequent updates to the branch targets corresponding to a cache line may occur due to the execution of instructions within the cache line. Instruction cache 16 provides an indication of the instruction address being fetched, so that branch prediction unit 14 may determine which branch target addresses to select for forming a branch prediction. Decode units 20 and functional units 24 provide update information to branch prediction unit 14. Because branch prediction unit 14 stores two targets per cache line, some branch instructions within the line may not be stored in branch prediction unit 14. Decode units 20 detect branch instructions which were not predicted by branch prediction unit 14. Functional units 24 execute the branch instructions and determine if the predicted branch direction is incorrect. The branch direction may be "taken", in which subsequent instructions are fetched from the target address of the branch instruction. Conversely, the branch direction may be "not taken", in which subsequent instructions are fetched from memory locations consecutive to the branch instruction. When a mispredicted branch instruction is detected, instructions subsequent to the mispredicted branch are discarded from the various units of microprocessor 10. A variety of suitable branch prediction algorithms may be employed by branch prediction unit 14.

Instructions fetched from instruction cache 16 are conveyed to instruction alignment unit 18. As instructions are fetched from instruction cache 16, the corresponding predecode data is scanned to provide information to instruction alignment unit 18 (and to MROM unit 34) regarding the instructions being fetched. Instruction alignment unit 18 utilizes the scanning data to align an instruction to each of decode units 20. In one embodiment, instruction alignment unit 18 aligns instructions from three sets of eight instruction bytes to decode units 20. Decode unit 20A receives an instruction which is prior to instructions concurrently received by decode units 20B and 20C (in program order). Similarly, decode unit 20B receives an instruction which is prior to the instruction concurrently received by decode unit 20C in program order. As used herein, the term "program order" refers to the order of the instruction as coded in the original sequence in memory. The program order of instructions is the order in which the instructions would be executed upon a microprocessor which fetches, decodes, executes, and writes the result of a particular instruction prior to fetching another instruction. Additionally, the term "dispatch" is used to refer to conveyance of an instruction to an issue position which is to execute the instruction. Issue positions may also dispatch load/store memory operations to load/store unit 26.

Decode units 20 are configured to decode instructions received from instruction alignment unit 18. Register operand information is detected and routed to register file 30 and reorder buffer 32. Additionally, if the instructions require one or more memory operations to be performed, decode

units 20 dispatch the memory operations to load/store unit 26. Each instruction is decoded into a set of control values for functional units 24, and these control values are dispatched to reservation stations 22 along with operand address information and displacement or immediate data which may be included with the instruction.

Microprocessor 10 supports out of order execution, and thus employs reorder buffer 32 to keep track of the original program sequence for register read and write operations, to implement register renaming, to allow for speculative instruction execution and branch misprediction recovery, and to facilitate precise exceptions. A temporary storage location within reorder buffer 32 is reserved upon decode of an instruction that involves the update of a register to thereby store speculative register states. If a branch prediction is incorrect, the results of speculatively-executed instructions along the mispredicted path can be invalidated in the buffer before they are written to register file 30. Similarly, if a particular instruction causes an exception, instructions subsequent to the particular instruction may be discarded. In this manner, exceptions are "precise" (i.e. instructions subsequent to the particular instruction causing the exception are not completed prior to the exception). It is noted that a particular instruction is speculatively executed if it is executed prior to instructions which precede the particular instruction in program order. Preceding instructions may be a branch instruction or an exception-causing instruction, in which case the speculative results may be discarded by reorder buffer 32.

The instruction control values and immediate or displacement data provided at the outputs of decode units 20 are routed directly to respective reservation stations 22. In one embodiment, each reservation station 22 is capable of holding instruction information (i.e., instruction control values as well as operand values, operand tags and/or immediate data) for up to three pending instructions awaiting issue to the corresponding functional unit. It is noted that for the embodiment of FIG. 1, each reservation station 22 is associated with a dedicated functional unit 24. Accordingly, three dedicated "issue positions" are formed by reservation stations 22 and functional units 24. In other words, issue position 0 is formed by reservation station 22A and functional unit 24A. Instructions aligned and dispatched to reservation station 22A are executed by functional unit 24A. Similarly, issue position 1 is formed by reservation station 22B and functional unit 24B; and issue position 2 is formed by reservation station 22C and functional unit 24C. As used herein, the term "issue position" refers to logic circuitry configured to receive an instruction and to execute that instruction. Once the instruction enters the issue position, it remains in that issue position until the execution of the instruction is completed.

Upon decode of a particular instruction, if a required operand is a register location, register address information is routed to reorder buffer 32 and register file 30 simultaneously. Those of skill in the art will appreciate that the x86 register file includes eight 32 bit real registers (i.e., typically referred to as EAX, EBX, ECX, EDX, EBP, ESI, EDI and ESP). In embodiments of microprocessor 10 which employ the x86 microprocessor architecture, register file 30 comprises storage locations for each of the 32 bit real registers. Additional storage locations may be included within register file 30 for use by MROM unit 34. Reorder buffer 32 contains temporary storage locations for results which change the contents of these registers to thereby allow out of order execution. A temporary storage location of reorder buffer 32 is reserved for each instruction which, upon decode, is

determined to modify the contents of one of the real registers. Therefore, at various points during execution of a particular program, reorder buffer 32 may have one or more locations which contain the speculatively executed contents of a given register. If following decode of a given instruction it is determined that reorder buffer 32 has a previous location or locations assigned to a register used as an operand in the given instruction, the reorder buffer 32 forwards to the corresponding reservation station either: 1) the value in the most recently assigned location, or 2) a tag for the most recently assigned location if the value has not yet been produced by the functional unit that will eventually execute the previous instruction. If reorder buffer 32 has a location reserved for a given register, the operand value (or reorder buffer tag) is provided from reorder buffer 32 rather than from register file 30. If there is no location reserved for a required register in reorder buffer 32, the value is taken directly from register file 30. If the operand corresponds to a memory location, the operand value is provided to the reservation station through load/store unit 26.

In one particular embodiment, reorder buffer 32 is configured to store and manipulate concurrently decoded instructions as a unit. This configuration will be referred to herein as "line-oriented". By manipulating several instructions together, the hardware employed within reorder buffer 32 may be simplified. For example, a line-oriented reorder buffer included in the present embodiment allocates storage sufficient for instruction information pertaining to three instructions (one from each decode unit 20) whenever one or more instructions are dispatched by decode units 20. By contrast, a variable amount of storage is allocated in conventional reorder buffers, dependent upon the number of instructions actually dispatched. A comparatively larger number of logic gates may be required to allocate the variable amount of storage. When each of the concurrently decoded instructions has executed, the instruction results are stored into register file 30 simultaneously. The storage is then free for allocation to another set of concurrently decoded instructions. Additionally, the amount of control logic circuitry employed per instruction is reduced because the control logic is amortized over several concurrently decoded instructions. A reorder buffer tag identifying a particular instruction may be divided into two fields: a line tag and an offset tag. The line tag identifies the set of concurrently decoded instructions including the particular instruction, and the offset tag identifies which instruction within the set corresponds to the particular instruction. It is noted that storing instruction results into register file 30 and freeing the corresponding storage is referred to as "retiring" the instructions. It is further noted that any reorder buffer configuration may be employed in various embodiments of microprocessor 10.

As noted earlier, reservation stations 22 store instructions until the instructions are executed by the corresponding functional unit 24. An instruction is selected for execution if: (i) the operands of the instruction have been provided; and (ii) the operands have not yet been provided for instructions which are within the same reservation station 22A-22C and which are prior to the instruction in program order. It is noted that when an instruction is executed by one of the functional units 24, the result of that instruction is passed directly to any reservation stations 22 that are waiting for that result at the same time the result is passed to update reorder buffer 32 (this technique is commonly referred to as "result forwarding"). An instruction may be selected for execution and passed to a functional unit 24A-24C during the clock cycle that the associated result is forwarded.

Reservation stations 22 route the forwarded result to the functional unit 24 in this case.

In one embodiment, each of the functional units 24 is configured to perform integer arithmetic operations of addition and subtraction, as well as shifts, rotates, logical operations, and branch operations. The operations are performed in response to the control values decoded for a particular instruction by decode units 20. It is noted that a floating point unit (not shown) may also be employed to accommodate floating point operations. The floating point unit may be operated similar to load/store unit 26 in that any of decode units 20 may dispatch instructions to the floating point unit. Additionally, functional units 24 may be configured to perform address generation for load and store memory operations performed by load/store unit 26.

Each of the functional units 24 also provides information regarding the execution of conditional branch instructions to the branch prediction unit 14. If a branch prediction was incorrect, branch prediction unit 14 flushes instructions subsequent to the mispredicted branch that have entered the instruction processing pipeline, and causes fetch of the required instructions from instruction cache 16 or main memory. It is noted that in such situations, results of instructions in the original program sequence which occur after the mispredicted branch instruction are discarded, including those which were speculatively executed and temporarily stored in load/store unit 26 and reorder buffer 32.

Results produced by functional units 24 are sent to reorder buffer 32 if a register value is being updated, and to load/store unit 26 if the contents of a memory location are changed. If the result is to be stored in a register, reorder buffer 32 stores the result in the location reserved for the value of the register when the instruction was decoded. A plurality of result buses 38 are included for forwarding of results from functional units 24 and load/store unit 26. Result buses 38 convey the result generated, as well as the reorder buffer tag identifying the instruction being executed.

Load/store unit 26 provides an interface between functional units 24 and data cache 28. In one embodiment, load/store unit 26 is configured with a load/store buffer having eight storage locations for data and address information for pending loads or stores. Decode units 20 arbitrate for access to the load/store unit 26. When the buffer is full, a decode unit must wait until load/store unit 26 has room for the pending load or store request information. Load/store unit 26 also performs dependency checking for load memory operations against pending store memory operations to ensure that data coherency is maintained. A memory operation is a transfer of data between microprocessor 10 and the main memory subsystem. Memory operations may be the result of an instruction which utilizes an operand stored in memory, or may be the result of a load/store instruction which causes the data transfer but no other operation. Additionally, load/store unit 26 may include a special register storage for special registers such as the segment registers and other registers related to the address translation mechanism defined by the x86 microprocessor architecture.

In one embodiment, load/store unit 26 is configured to perform load memory operations speculatively. Store memory operations are performed in program order, but may be speculatively stored into the predicted way. If the predicted way is incorrect, the data prior to the store memory operation is subsequently restored to the predicted way and the store memory operation is performed to the correct way. In another embodiment, stores may be executed specula-

13

tively as well. Speculatively executed stores are placed into a store buffer, along with a copy of the cache line prior to the update. If the speculatively executed store is later discarded due to branch misprediction or exception, the cache line may be restored to the value stored in the buffer. It is noted that load/store unit 26 may be configured to perform any amount of speculative execution, including no speculative execution.

Data cache 28 is a high speed cache memory provided to temporarily store data being transferred between load/store unit 26 and the main memory subsystem. In one embodiment, data cache 28 has a capacity of storing up to sixteen kilobytes of data in an eight way set associative structure. Similar to instruction cache 16, data cache 28 may employ a way prediction mechanism. It is understood that data cache 28 may be implemented in a variety of specific memory configurations, including a set associative configuration.

In one particular embodiment of microprocessor 10 employing the x86 microprocessor architecture, instruction cache 16 and data cache 28 are linearly addressed. The linear address is formed from the offset specified by the instruction and the base address specified by the segment portion of the x86 address translation mechanism. Linear addresses may optionally be translated to physical addresses for accessing a main memory. The linear to physical translation is specified by the paging portion of the x86 address translation mechanism. It is noted that a linear addressed cache stores linear address tags. A set of physical tags (not shown) may be employed for mapping the linear addresses to physical addresses and for detecting translation aliases. Additionally, the physical tag block may perform linear to physical address translation.

Turning now to FIG. 2, a block diagram of one embodiment of decode units 20B and 20C are shown. Each decode unit 20 receives an instruction from instruction alignment unit 18. Additionally, MROM unit 34 is coupled to each decode unit 20 for dispatching fast path instructions corresponding to a particular MROM instruction. Decode unit 20B comprises early decode unit 40B, multiplexor 42B, and opcode decode unit 44B. Similarly, decode unit 20C includes early decode unit 40C, multiplexor 42C, and opcode decode unit 44C.

Certain instructions in the x86 instruction set are both fairly complicated and frequently used. In one embodiment of microprocessor 10, such instructions include more complex operations than the hardware included within a particular functional unit 24A-24C is configured to perform. Some of such instructions are classified as a special type of MROM instruction referred to as a "double dispatch" instruction. These instructions are dispatched to a pair of opcode decode units 44 by MROM unit 34. It is noted that opcode decode units 44 are coupled to respective reservation stations 22. Each of opcode decode units 44A-44C forms an issue position with the corresponding reservation station 22A-22C and functional unit 24A-24C. Instructions are passed from an opcode decode unit 44 to the corresponding reservation station 22 and further to the corresponding functional unit 24.

Multiplexor 42B is included for selecting between the instructions provided by MROM unit 34 and by early decode unit 40B. During times in which MROM unit 34 is dispatching instructions, multiplexor 42B selects instructions provided by MROM unit 34. At other times, multiplexor 42B selects instructions provided by early decode unit 40B. Similarly, multiplexor 42C selects between

14

instructions provided by MROM unit 34, early decode unit 40B, and early decode unit 40C. The instruction from MROM unit 34 is selected during times in which MROM unit 34 is dispatching instructions. During times in which early decode unit 40A detects a double dispatch instruction, the instruction from early decode unit 40B is selected by multiplexor 42C. Otherwise, the instruction from early decode unit 40C is selected. Selecting the instruction from early decode unit 40B into opcode decode unit 44C allows a fast path instruction decoded by decode unit 20B to be dispatched concurrently with a double dispatch instruction decoded by decode unit 20A. In this manner, instruction alignment unit 18 need not attempt to align MROM instructions and concurrently dispatched fast path instructions to their final issue positions. Instead, the instructions may be aligned to a position and then adjusted between early decode units 40 and opcode decode units 44.

According to one embodiment employing the x86 instruction set, early decode units 40 perform the following operations:

- (i) merge the prefix bytes of the instruction into an encoded prefix byte;
- (ii) decode unconditional branch instructions (which may include the unconditional jump, the CALL, and the RETURN) which were not detected during branch prediction;
- (iii) decode source and destination flags;
- (iv) decode the source and destination operands which are register operands and generate operand size information; and
- (v) determine the displacement and/or immediate size so that displacement and immediate data may be routed to the opcode decode unit.

Opcode decode units 44 are configured to decode the opcode of the instruction, producing control values for functional unit 24. Displacement and immediate data are routed with the control values to reservation stations 22.

Since early decode units 40 detect operands, the outputs of multiplexors 42 are routed to register file 30 and reorder buffer 32. Operand values or tags may thereby be routed to reservation stations 22. Additionally, memory operands are detected by early decode units 40. Therefore, the outputs of multiplexors 42 are routed to load/store unit 26. Memory operations corresponding to instructions having memory operands are stored by load/store unit 26.

Turning now to FIG. 3, a diagram depicting instruction processing pipeline stages for one embodiment of microprocessor 10 is shown. Other embodiments of microprocessor 10 may employ dissimilar instruction processing pipelines. The instruction processing pipeline shown in FIG. 3 includes an instruction fetch stage 50, an instruction scan stage 52, a first alignment stage 54, a second alignment stage 56, an early decode stage 58, a decode stage 60, an MROM entry point stage 62, an MROM access stage 64, and an MROM early decode stage 66. MROM entry point stage 62, MROM access stage 64, and MROM early decode stage 66 correspond to MROM unit 34. Instruction fetch stage 50 and instruction scan stage 52 are performed by instruction cache 16. Similarly, first and second alignment stages 54 and 56 correspond to instruction alignment unit 18, early decode stage 58 corresponds to early decode units 40, and decode stage 60 corresponds to opcode decode units 44.

During instruction fetch stage 50, instructions are fetched from instruction cache 16. The instruction cache storage is accessed via a fetch address provided by branch prediction unit 14, and instructions are conveyed to an instruction

scanning unit within microprocessor 10. During instruction scan stage 52, the instructions are scanned and instruction blocks are created. The instruction blocks are conveyed to the byte queue within first alignment stage 54. Additionally, MROM instructions are detected during instruction scan stage 52. A detected MROM instruction is routed to MROM entry point stage 62. In one embodiment, MROM unit 34 is configured to accept one instruction per clock cycle. Therefore, if a second MROM instruction is detected within a set of instruction bytes being scanned during a particular clock cycle, instruction blocks including the second MROM instruction and subsequent instructions in program order are stalled until a subsequent clock cycle.

During first alignment stage 54, instructions are selected from the byte queue included therein for dispatch. In one embodiment, up to four instructions are selected for dispatch from which up to three instructions are actually dispatched, as detailed further below. The instructions are conveyed to second alignment stage 56, subsequently to early decode stage 58 and then to decode stage 60.

MROM entry point stage 62 is used to determine a location within the ROM storage of MROM unit 34 at which the first instructions corresponding to a particular MROM instruction are stored. The address is passed to MROM access stage 64, which accesses the ROM storage and receives the instructions stored therein. In one embodiment, a line of instructions (i.e. up to the number of instructions which may be stored in a reorder buffer line) are received during one cycle of ROM storage access. The line of instructions is then transmitted to MROM early decode stage 66, which formats the instructions similar to the formatting of early decode units 40 (such that opcode decode units 44 detect only one type of instruction formatting). The line of instructions is then inserted into opcode decode units 44 via multiplexors 42. For MROM instructions which employ more than a single line of instructions, additional MROM accesses are performed in MROM access stage 64 and subsequent lines of instructions conveyed to MROM early decode stage 66 during subsequent clock cycles. First alignment stage 54 and second alignment stage 56 are stalled during such subsequent clock cycles.

It is noted that, since first alignment stage 54 includes a byte queue storing multiple instructions and the instructions are selected therefrom in program order, a particular MROM instruction may arrive in MROM access stage 64 prior to being selected for dispatch from the byte queue. The particular MROM instruction may be subsequent to a large number of instructions within the byte queue, and instructions are selected for dispatch in program order. (MROM instructions are routed to MROM unit 34 but are not removed from the instruction blocks conveyed to instruction alignment unit 18.) Alternatively, the particular MROM instruction may be queued in MROM unit 34 while a prior MROM instruction executes. The particular MROM instruction may be selected for dispatch prior to arriving at MROM access stage 64. Therefore, synchronization is provided between second alignment stage 56 and MROM access stage 64 (illustrated by synchronization bus 67).

When MROM access stage 64 receives an entry point address from MROM entry point stage 62, MROM access stage 64 informs second alignment stage 56 by asserting a signal upon synchronization bus 67. When second alignment stage 56 receives a dispatched MROM instruction from first alignment stage 54, second alignment stage 56 signals MROM access stage 56 via synchronization bus 67. In this manner, the MROM instruction progresses to both MROM early decode stage 66 and early decode stage 58 during the

same clock cycle. Because both second alignment stage 56 and MROM access stage 64 receive instructions in program order, it is sufficient to synchronize instructions via synchronization bus 67.

During MROM entry point stage 62, MROM unit 34 determines if a particular MROM instruction is double dispatch. A particular MROM instruction is double dispatch if the particular MROM instruction corresponds to a single line of instructions within which two instructions are stored. If MROM unit 34 detects a double dispatch instruction, a double dispatch signal upon a double dispatch conductor 68 is asserted. Otherwise, the double dispatch signal is deasserted. The double dispatch signal is conveyed to both first alignment stage 54 and second alignment stage 56. Second alignment stage 56 uses the state of the double dispatch signal to determine if instructions dispatched during the previous clock cycle (in first alignment stage 54 and therefore currently residing in second alignment stage 56) should be discarded. More particularly, second alignment stage 56 discards the second instruction in program order if: (i) an MROM instruction and a fast path instruction were concurrently dispatched; and (ii) the double dispatch signal is deasserted. Otherwise, second alignment stage 56 passes the instructions to early decode stage 58. It is noted that, although MROM entry point stage 62 detects the double dispatch nature of an MROM instruction, the double dispatch signal as shown in FIG. 3 is asserted from the MROM access stage 64. Alternatively, the double dispatch signal may be asserted from MROM entry point stage 62 and instruction alignment unit 18 may store the signal value for use in the subsequent clock cycle.

First alignment stage 54 uses the double dispatch signal as well. When first alignment stage 54 concurrently dispatches an MROM instruction and a fast path instruction (referred to herein as "packing"), the second of the two instructions in program order is retained in the byte queue. During each clock cycle, first alignment stage 54 initially selects up to four instructions for dispatch during a particular clock cycle. If first alignment stage 54 packed during the previous clock cycle and the double dispatch signal is asserted, then the first of the four instructions (in program order) is ignored and the remainder are dispatched. Conversely, if first alignment stage 54 did not pack during the previous clock cycle or the double dispatch signal is deasserted, the first three of the four instructions (in program order) are dispatched and the fourth is retained by the byte queue. In this manner, redispach of the second of the packed instructions is performed when needed without sacrificing other dispatch positions.

Turning now to FIG. 4, a block diagram of one embodiment of instruction cache 16 and instruction alignment unit 18 are shown. Instruction cache 16 includes an instruction cache storage and control block 70 and an instruction scanning unit 72. Instruction alignment unit 18 includes a byte queue 74, a selection control unit 76, and a multiplex to issue block 78.

Instruction cache storage and control block 70 includes storage for instruction cache lines and related control circuitry for fetching instructions from the storage, for selecting cache lines to discard when a cache miss is detected, etc. Instruction cache storage and control block 70 receives fetch addresses from branch prediction unit 14 (not shown) in order to fetch instructions for execution by microprocessor 10. Instruction bytes fetched from instruction cache storage and control block 70 are conveyed to instruction scanning unit 72 upon an instructions bus 80. Instruction bytes are conveyed upon instructions bus 80, as well as corresponding predecode data (e.g. start, end, and functional bits). In one

embodiment, sixteen bytes stored in contiguous memory locations are conveyed upon instructions bus 80 along with the corresponding predecode data. The sixteen bytes form either the upper or lower half of the 32 byte cache line. The upper half of the cache line is the half stored in memory addresses having larger numerical values, while the lower half is stored in memory addresses having smaller numerical values. Additionally, instruction scanning unit 72 receives information regarding the bytes within the sixteen bytes which are to be conveyed as instructions to instruction alignment unit 18. Instruction bytes at the beginning of the sixteen bytes may be ignored if the bytes are fetched as the target of a branch instruction, and the target address identifies a byte other than the first byte of the sixteen bytes. Additionally, if a branch instruction is within the sixteen bytes and branch prediction unit 14 predicts the branch taken, then bytes subsequent to the branch instruction within the sixteen bytes are ignored.

Instruction scanning unit 72 scans the predecode data associated with the bytes which are to be conveyed as instructions to instruction alignment unit 18. In the present embodiment, instruction scanning unit 72 divides the sixteen bytes conveyed by instruction cache storage and control block 70 into two portions comprising eight contiguous bytes each. One portion forms the lower half of the sixteen bytes (i.e. the bytes stored at smaller numerical addresses than the bytes forming the upper half of the sixteen bytes). The other portion forms the upper half of the sixteen bytes. Therefore, an eight byte portion forms one of four quarters of the 32 byte cache line employed by instruction cache storage and control block 70, according to the present embodiment. As used herein, bytes are contiguous if they are stored in contiguous memory locations in the main memory subsystem. It is noted that particular sizes of various components, such as instruction block sizes, are used herein for clarity of the description. Any size may be used for each component within the spirit and scope of the appended claims.

Instruction scanning unit 72 scans the predecode data of each portion of the instructions independently and in parallel. Instruction scanning unit 72 identifies up to a predefined maximum number of instructions within each portion from the start and end byte information included within the predecode data. For the present embodiment, the predefined maximum number is three. Generally speaking, instruction scanning unit 72 preferably identifies a maximum number of instructions in each portion equal to the number of issue positions included within microprocessor 10.

The instruction bytes and instruction identification information generated by instruction scanning unit 72 are conveyed to byte queue 74 upon an instructions bus 82 and an instruction data bus 84, respectively. The instruction bytes are conveyed as eight byte portions, and the instruction data is arranged accordingly such that each eight byte portion is associated with a portion of the instruction identification information conveyed upon instruction data bus 84. Each eight byte portion and the corresponding instruction identification information forms an instruction block. It is noted that, although an instruction block includes eight bytes in the present embodiment, instruction blocks may include any number of bytes in various embodiments. Byte queue 74 receives the instruction blocks conveyed and stores them into one of multiple subqueues included therein. In the embodiment shown, byte queue 74 includes three subqueues: a first subqueue 86A, a second subqueue 86B, and a third subqueue 86C. First subqueue 86A stores the instruction block which is foremost among the instruction blocks

stored in byte queue 74 in program order. Second subqueue 86B stores the instruction block which is second in program order, and third subqueue stores the instruction block which is third in program order. It is noted that various embodiments of byte queue 74 may include any number of subqueues 66.

If a particular portion as scanned by instruction scanning unit 72 includes more than the maximum predefined number of instructions, then the particular portion is retained by instruction scanning unit 72. During the following clock cycle, the particular eight byte portion is scanned again. The predecode data corresponding to the previously identified instructions is invalidated such that instruction scanning unit 72 detects the additional instructions. If the other portion concurrently received with the particular portion is subsequent to the particular portion in program order, then the other portion is rescanned as well. Byte queue 74 discards the instruction blocks received from the other portion, in order to retain program order among the instruction blocks stored in the byte queue.

A control unit 90 within byte queue 74 conveys a byte queue status upon byte queue status bus 88 to instruction scanning unit 72. Byte queue status bus 88 includes a signal corresponding to each subqueue 86. The signal is asserted if the subqueue 86 is storing an instruction block, and deasserted if the subqueue 86 is not storing an instruction block. In this manner, instruction scanning unit 72 may determine how many instruction blocks are accepted by byte queue 74 during a clock cycle. If two instruction blocks are conveyed during a clock cycle and only one instruction block is accepted, instruction scanning unit 72 retains the rejected instruction block and rescans the instruction block in the subsequent clock cycle.

As noted above, an instruction block may contain up to a predefined maximum number of instructions (e.g. three in the present embodiment). Additionally, eight contiguous bytes are conveyed for each instruction block in the present embodiment. However, due to the variable byte length of the x86 instructions, an instruction may begin within one set of contiguous bytes and end in another set of contiguous bytes, referred to as overflow instructions. If an overflow instruction is detected, it is identified as the last of the predefined number of instructions. Instead of being indicated as a valid instruction within the instruction block, the overflow instruction is identified as an overflow. Instruction identification information is generated, but the instruction is handled somewhat differently, as will be explained in more detail below.

In one embodiment, the instruction identification information for each instruction includes: (i) start and end pointers identifying the bytes at which the identified instruction begins and ends within the eight bytes; (ii) a valid mask containing eight bits, one for each of the eight bytes; (iii) a bit indicative of whether the instruction is MROM or fast path; and (iv) an instruction valid bit indicating that the instruction is valid and an overflow bit for the last instruction indicating that it is an overflow. The valid mask includes a binary one bit corresponding to each byte included within the particular instruction (i.e. the bits between the start pointer and end pointer, inclusive, are set). Zero bits are included for the other bytes. Additional information conveyed with the instruction identification information is the taken/not taken prediction if the instruction is a branch instruction, bits indicating which of the quarters of the 32 byte cache line the eight bytes correspond to, the functional bits from the predecode data corresponding to the eight bytes, and a segment limit identifying the segment limit

19

within the eight bytes for exception handling. The additional information is provided by instruction cache storage and control block 70 except for the branch prediction, which is provided by branch prediction unit 14.

Selection control unit 76 examines the instruction identification information stored in each subqueue to generate selection controls for multiplex to issue block 78. Multiplex to issue block 78 includes a plurality of multiplexors for selecting instruction bytes from byte queue 74 for conveyance to each of decode units 20. Byte queue 74 maintains certain properties with respect to each subqueue 86 in order to simplify the selection logic within selection control unit 76, as will be explained in more detail below. Instructions are selected and conveyed, and corresponding instruction identification information is invalidated such that subsequent instructions may be dispatched in subsequent clock cycles.

Subqueues 86 store instruction information in a plurality of instruction positions (or simply "positions"). The number of instruction positions is preferably equal to the maximum number of instructions which may be included in an instruction block. For the present embodiment, three positions are included. The first position ("position 10") stores the instruction identification information corresponding to the instruction which is foremost in program order within the instruction block stored in the subqueue 86. The second position ("position 11") stores the instruction identification information corresponding to the second instruction in program order within the instruction block. Finally, the third position ("position 12") stores the instruction identification information corresponding to the last instruction in program order. Alternatively, position 12 may store instruction identification information corresponding to an overflow instruction. Certain instruction identification information is the same for each instruction (e.g. the segment limit). To avoid duplicating information, this instruction information may be stored as a single copy separate from the instructions positions.

Control unit 90 maintains the information stored in each subqueue 86. In particular, control unit 90 directs each subqueue 86 to shift instruction identification information between the positions when instructions are selected for dispatch. For example, if the instruction corresponding to position 10 is dispatched, the information stored in position 11 is shifted into position 10 and the information stored in position 12 is shifted into position 11. Similarly, if the instructions corresponding to positions 10 and 11 are dispatched, then information stored in position 12 is shifted into position 10. In this manner, the instruction within the subqueue which is foremost in program order is maintained in position 10, the instruction which is second in program order is maintained in position 11, etc. In order to select an instruction for dispatch to decode unit 20A, selection control unit 76 examines the instruction identification information stored in position 10 of each subqueue. Advantageously, a small amount of logic may be employed to select the instruction. Similarly, position 10 of subqueue 86A and position 12 of each subqueue 86A-86C are not examined to select an instruction for decode unit 20B. The second instruction to be dispatched will be found within the first two positions of one of the subqueues 66 when maintained in accordance with the above. Selection control unit 76 informs control unit 90 of which instructions positions were selected for dispatch during a clock cycle, such that subqueue shifting may be performed.

According to one embodiment, instruction identification information is shifted internally to each subqueue 86 independently. Instruction identification information is not,

20

therefore, shifted from position 10 of subqueue 86B into positions within subqueue 86A. Instead, when each of the instructions within subqueue 86A have been dispatched, subqueue 86B is shifted into subqueue 86A as a whole. The logic for shifting between subqueues 86 may operate independently from and in parallel with the internal shifting of each subqueue 86A-86C.

Position 12 may store instruction identification information regarding an overflow instruction. If position 12 is storing information regarding an overflow instruction, then the information is not shifted to position 10 or 11 as described above. In this manner, overflow instruction information is always available in position 12. Selection control unit 76 may examine the information stored in position 12 for routing bytes corresponding to an overflow instruction, as opposed to having to locate the overflow information within the positions and then determining byte routing.

Selection control unit 76 selects instructions from the instructions positions within subqueues 86 for potential dispatch. The instructions selected are the instructions which are foremost in program order among the instructions stored in subqueues 86. More instructions are initially selected for dispatch than the number of issue positions included in microprocessor 10, in order to correctly perform redispach of instructions when an MROM instruction and a fast path instruction are concurrently dispatched and the MROM instruction is found to be an arbitrary dispatch instruction. Selection control unit 76 then selects from the potentially dispatchable instructions based upon the value of a packed state stored in a packed state register 92 coupled to selection control unit 76 and the state of the double dispatch signal upon double dispatch conductor 68, also coupled to selection control unit 76.

When selection control unit 76 selects an MROM instruction and a fast path instruction for concurrent dispatch during a clock cycle, selection control unit 76 sets the packed state. Otherwise, the packed state is reset. The packed state so generated is stored into packed state register 92 for use during the succeeding clock cycle. Additionally, selection control unit 76 informs control unit 90 that the first of the MROM instruction and the fast path instruction (in program order) is being dispatched. In this manner, byte queue 54 retains the second of the two instructions in program order, despite the dispatch of the second of the two instructions. In one embodiment, the packed state comprises a bit indicative, when set, that an MROM instruction and a fast path instruction were concurrently dispatched in the previous clock cycle.

From the potentially dispatchable instructions, selection control unit 76 selects instructions for dispatch based upon the packed state stored in packed state register 92 and the double dispatch signal. If the packed state is set, an MROM instruction and a fast path instruction were concurrently dispatched in the previous clock cycle. Therefore, the instruction within the potentially dispatchable instructions which is foremost in program order is one of the two instructions previously dispatched when the packed state is set. If the packed state is set and the double dispatch signal is asserted, the concurrent dispatch of the MROM instruction and the fast path instruction is successful. If the packed state is set and the double dispatch signal is deasserted, the concurrent dispatch of the MROM instruction and the fast path instruction is unsuccessful. The MROM instruction occupies at least three issue positions, and therefore the fast path instruction cannot be concurrently dispatched for the embodiment of microprocessor 10 shown in FIG. 1. If the packed state is clear, concurrent dispatch of an MROM and

fast path instructions was not performed in the previous clock cycle. Therefore, the instructions within the potentially dispatchable instructions were not previously dispatched.

According to one embodiment, selection control unit 76 selects the foremost instructions in program order from the set of potentially dispatchable instructions if either the packed state is clear or the packed state is set and the double dispatch signal is deasserted. In the case of the packed state being clear, the foremost set of instructions are dispatched and program order is maintained. In the case of the packed state being set and the double dispatch signal being deasserted, the second of the instructions dispatched during the previous clock cycle is redispached. If the second of the instructions is the MROM instruction, it is dispatched alone. If the second of the instructions is the fast path instruction, additional instructions may be selected for concurrent dispatch. Advantageously, the largest number of concurrently dispatchable instructions is selected, even in the case of redispaching a previously dispatched instruction.

If the packed state is set and the double dispatch signal is asserted, then the instruction within the potentially dispatched instructions which is foremost in program order is the second of the previously dispatched instructions and that instruction is successfully dispatched during the previous clock cycle (i.e. the MROM instruction and fast path instruction, when taken together, occupy a number of issue positions less than or equal to the number of issue positions included within microprocessor 10). This instruction is therefore not selected during the current clock cycle. Instead, instructions are dispatched from the remaining of the potentially dispatchable instructions.

Upon selection of the instructions dispatched, the packed state is determined for the subsequent clock cycle. In addition, control unit 90 is informed of the instructions dispatched. For the case of the packed state being set and the double dispatch signal being asserted, the instruction which was previously dispatched is indicated as dispatched as well as each of the instructions dispatched during the present clock cycle. Subqueues 86 are shifted accordingly. In one embodiment, control unit 90 is informed of the subqueue and position storing the last instruction (in program order) to be dispatched. Selection control unit 76 identifies the last instruction in accordance with the above functionality. Byte queue 74 shifts out the instructions prior to and including the indicated last instruction. In this manner, byte queue 74 operates independent of the logic used to concurrently dispatch MROM and fast path instructions. For example, when packing an MROM instruction and a fast path instruction, the first of the instructions in program order is marked as the last instruction. The second of the instructions is thereby retained in byte queue 74 while the first of the instructions is shifted out.

It is noted that, in one embodiment, the circuitry shown in FIG. 4 for instruction alignment unit 18 forms first alignment stage 54. Second alignment stage 56 is not shown in FIG. 4. It is further noted that additional details regarding the operation of byte queue 74 may be found in the commonly assigned, co-pending patent application entitled: "A Byte Queue Divided into Multiple Subqueues for Optimizing Instruction Selection Logic", filed concurrently herewith by Narayan, et al. The disclosure of the referenced patent application is incorporated herein by S reference in its entirety.

Turning next to FIG. 5, a flowchart 100 depicting the operation of selection control unit 76 is shown according to one exemplary embodiment. During step 102, selection

control unit 76 selects a set of potentially dispatchable instructions. For the present embodiment, up to four instructions (e.g. instructions A, B, C, and D, in program order) may be selected. Preferably, the maximum number of instructions selected into the set of potentially dispatchable instructions is the number of issue positions included in microprocessor 10 plus the number of instructions which may be redispached due to an unsuccessful concurrent dispatch of MROM instructions and fast path instructions during a previous clock cycle. Therefore, the maximum number of instructions within the potentially dispatchable instructions may vary from embodiment to embodiment.

The selection of instructions according to step 102 may involve certain restrictions. For example, the present embodiment may concurrently dispatch up to three fast path instructions (one for each issue position), a fast path instruction and an MROM instruction, or an MROM instruction alone. Therefore, instructions C and D may not be MROM instructions in the present embodiment. If the instruction which would otherwise be instruction C is an MROM instruction, no instructions are selected as instruction C or D. Furthermore, microprocessor 10 allows up to one predicted taken branch instruction to be concurrently dispatched, according to one embodiment. If a second predicted taken branch instruction is encountered, selection control unit 76 does not select that branch instruction or any subsequent instructions for potential dispatch. According to another embodiment, instructions from at most two cache lines may be concurrently dispatched. If an instruction from a third cache line is encountered, it is not selected for concurrent dispatch. These restrictions may not be applied in other embodiments. Additional or supplemental restrictions may be applied in other embodiments as well.

Decision box 104 determines which of the set of potentially dispatchable instructions are selected for dispatch, based upon the packed state and the double dispatch signal. If the packed state is set and the double dispatch signal is asserted, instructions B, C, and D are selected for dispatch (step 106). In this case, instruction A is one of the previously dispatched instructions. Since the concurrent dispatch of the MROM instruction and the fast path instruction is successful, instruction A need not be redispached during the present clock cycle. If the packed state is clear or the double dispatch signal is deasserted, then instructions A, B, and C are selected for dispatch (step 108).

Following selection of instructions for dispatch, selection control unit 76 determines the packed state for the subsequent clock cycle (decision block 110). For the present embodiment, if the selected instructions include an MROM instruction and a fast path instruction, then the packed state is set (step 112). The packed state is clear if the selected instructions do not include an MROM instruction and a fast path instruction (step 114).

Generally speaking, a selection method similar to flowchart 100 may be used to speculatively dispatch a set of instructions concurrently. The dispatch is speculative in the sense that the dispatched set of instructions may subsequently be determined to occupy a number of issue positions greater than the number of issue positions included within the microprocessor. Upon such speculative dispatch, the packed state may be set and a number of instructions equal to the maximum number of instructions which may be redispached are retained in the byte queue. The retained instructions as well as a set of additional instructions may then be preliminarily selected as potentially dispatchable instruction in the succeeding clock cycle, and appropriate selection of instructions from the potentially dispatchable

instructions may be performed based upon the result of the speculative dispatch.

Turning next to FIG. 6, a logic diagram of one embodiment of multiplex to issue block 78 is shown. Multiplex to issue block 78 includes a plurality of first multiplexors 120A-120D and a plurality of second multiplexors 122A-122C. First multiplexors 120 receive a set of selection controls upon a first selection controls bus 124 from selection control unit 76. Each first multiplexor 120 receives a separate set of selection controls, in one embodiment. Similarly, a selection control for multiplexors 122 are received upon second selection controls bus 126. In one embodiment, one selection control is included upon selection controls bus 126. The selection control is shared by second multiplexors 122. Second multiplexor 122A is coupled to first multiplexors 120A and 120B. Similarly, each second multiplexor 122 is coupled to a pair of first multiplexors 120 as shown in FIG. 6. Additionally, each of second multiplexors 122 selects an instruction for conveyance to a corresponding decode unit 20. However, the instruction selected may pass through additional instruction processing pipeline stages prior to arrival in the corresponding decode unit. For example, one embodiment of microprocessor 10 employs the instruction processing pipeline shown in FIG. 3. Instructions pass from second multiplexors 122 through second alignment stage 56 prior to arrival in decode units 20 in that embodiment.

Selection control unit 76 generates selection controls for each first multiplexor 120 by scanning the position information stored in byte queue 74. First multiplexor 120A produces the instruction which is foremost in program order within byte queue 74. Similarly, multiplexors 120B, 120C, and 120D produce the second, third, and fourth instructions in program order from the instructions stored in byte queue 74, respectively.

Second multiplexors 122 are provided for selecting instructions from the potentially dispatchable instructions identified by first multiplexors 120. The selection control upon second selection control bus 126 is toggled in conformance with the flowchart shown in FIG. 5. In other words, the selection control is toggled to cause second multiplexor 122A to select the output of first multiplexor 120A, second multiplexor 122B to select the output of first multiplexor 120B, and second multiplexor 122C to select the output of first multiplexor 120C if step 108 is performed. Alternatively, when step 106 is performed the selection control is toggled to cause second multiplexor 122A to select the output of first multiplexor 120B, second multiplexor 122B to select the output of first multiplexor 120C, and second multiplexor 122C to select the output of first multiplexor 120D.

Turning next to FIG. 7, a table 130 is shown identifying the valid combinations of instructions which may be stored within a particular subqueue according to the present embodiment. Other embodiments may employ similar or dissimilar combinations of instructions. Each row of table 130 is a valid combination of instructions, listed by position (10-12 as shown across the top of table 130). The symbol "X" in a position indicates an invalid instruction (i.e. no instruction identification information is stored therein). The symbols "A", "B", and "C" indicate that valid instruction identification information is stored in that position. Symbol "A" identifies an instruction prior to instructions identified by symbols "B" and "C" in program order. Similarly, symbol "B" identifies an instruction prior to the instruction identified by symbol "C" in program order. The symbol "O" in position 12 indicates that instruction identification information

is stored therein with respect to an overflow instruction. The symbol "O" represents an instruction subsequent to the instructions identified by symbols "A" and "B" in program order.

Table 130 illustrates that instruction identification information is stored in position 11 only when instruction identification information is stored in position 10. Similarly, instruction identification information is stored in position 12 only when instruction identification information is stored in both positions 10 and 11, except when position 12 stores an overflow instruction. When position 12 stores an overflow instruction, position 11 still stores instruction identification information only if position 10 stores instruction identification information. However, position 12 stores the overflow instruction information independent of the status of positions 10 and 11. Advantageously, the instruction which is foremost in program order is stored in position 10, even after instructions from the instruction block have been dispatched. Additionally, overflow instructions remain stored in position 12, even after instructions from the instruction block have been dispatched.

Turning next to FIG. 8, a table 132 is shown depicting the positions analyzed by selection control unit 76 for selecting instructions via each first multiplexor 120 in accordance with one embodiment. Each row of table 132 corresponds to a particular first multiplexor 120A-120D, as identified by the first column of table 132. For first multiplexor 120A, selection control unit 76 analyzes the instruction identification information stored in position 10 of each subqueue 86. Additionally, the overflow bit is examined for subqueues 86A and 86B. If position 10 of first subqueue 86A is valid, then that instruction is selected via multiplexor 120A. If position 10 of second subqueue 86B is valid and position 10 of first subqueue 86A is invalid, then the instruction corresponding to position 10 of second subqueue 86B is selected. Finally, the instruction corresponding to position 10 of third subqueue 86C is selected if position 10 of subqueues 86A and 86B are invalid. Because instructions may be up to 15 bytes long, an instruction may begin in first subqueue 86A, overflow into second subqueue 86B, and further overflow into third subqueue 86C. Such a case is an example where position 10 of third subqueue 86C is selected. In addition, the overflow indications from first and second subqueues 86A and 86B are considered in creating multiplexer selection controls for first multiplexor 120A. If position 10 of second subqueue 86B is selected and the overflow indication of first subqueue 86A indicates overflow, then instruction bytes from first subqueue 86A form a portion of the instruction indicated by position 10 of second subqueue 86B. The start pointer and valid mask corresponding to position 12 of subqueue 86A are used to multiplex the instruction bytes with instruction bytes indicated by the end pointer and valid mask corresponding to instruction 10 of subqueue 86B (or subqueue 86C, if the instruction overflows thereto).

For first multiplexor 120B, position 11 of first subqueue 86A is analyzed along with positions 10 and 11 of second and third subqueues 86B and 86C. Because byte queue 74 maintains each subqueue such that instructions are shifted to occupy positions 10 and 11 when previous instructions within the subqueue are dispatched, position 11 is the only position within first subqueue 86A which may contain the second instruction in program order which is to be dispatched during a clock cycle. Similarly, positions 10 and 11 of subqueues 86B and 86C may contain the second instruction to be dispatched. If a position 10 is selected for dispatch, the overflow indication of the preceding subqueue is also analyzed for forming multiplexor selection controls.

For first multiplexor 120C, position 12 of first subqueue 86A is analyzed along with each of the positions of second and third subqueues 86B and 86C. Similar to the above discussion, if position 10 of a subqueue is selected, the overflow indication of the preceding subqueue is analyzed to determine multiplexor selection controls. Finally, selection control unit 76 considers the positions of subqueues 86B and 86C to determine selection controls for first multiplexor 120D. Since first multiplexor 120D selects the fourth instruction in program order, positions within first subqueue 86A are not considered for selection via multiplexor 120D. Positions within first subqueue 86A store at most the first three instructions in program order, according to the present embodiment.

Table 132 illustrates certain advantages of operating byte queue 74 in the manner described herein. Selection of the instructions for first multiplexor 120A involves analyzing only three positions out of the nine positions included within byte queue 74. If byte queue 74 were implemented as, for example, a circular buffer, then each of the positions would have to be considered for dispatch to issue position zero. Similarly, analysis of only a few issue positions is performed to select the instruction for first multiplexor 120B. Selection logic is thereby reduced, allowing for fewer cascaded levels of logic. A higher operating frequency for microprocessor 10 may thereby be achieved.

Table 132 shows the positions analyzed to select instructions for issue. However, even though an instruction may be selectable for issue based upon table 132, other factors may cause a particular instruction not to be issued. For example, the instruction in position 10 of second subqueue 86B may be selected by multiplexor 120B according to table 132. However, the instruction in position 10 of first subqueue 86A may be an MROM instruction. Because microprocessor 10 dispatches MROM instructions without concurrent issue of other MROM instructions, the instruction in position 10 of second subqueue 86B is not selected if it is an MROM instruction. Other such restrictions may be imposed depending upon the embodiment of microprocessor 10, and may be included within the logic of selection control unit 58.

Turning next to FIG. 9, a table 134 is shown depicting the allowable issue position combinations (i.e. the selections by second multiplexors 122 under control of selection control unit 76) according to one embodiment of microprocessor 10. Each row of table 134 indicates an allowable combination of instructions, wherein each column is an issue position. A "-" in an issue position indicates that no instruction is dispatched to that issue position. An instruction may not be dispatched to a particular issue position for a variety of reasons. For example, dispatch restrictions included according to a particular embodiment of microprocessor 10 may cause an issue position not to receive an instruction during a clock cycle. One embodiment of microprocessor 10 restricts concurrent issue of instructions with an MROM instruction to up to one fast path instruction. Therefore, rows 136 and 138 are allowable combinations in which issue position two does not receive an instruction. Additionally, byte queue 74 may contain only a few instructions during a particular clock cycle (in the case of a fetch miss in instruction cache 16, for example). Therefore, selection control unit 76 selects the available instructions and no instructions are conveyed in the remaining issue positions. An "F" in an issue position indicates a fast path instruction, while an "M" indicates an MROM instruction.

Rows 136 and 138 are cases in which the packed state described above is set, for embodiments employing the set of allowable issue position combinations represented by

table 134. Other rows in table 134 do not cause the setting of the packed state.

Turning now to FIG. 10, a first example of operation of selection control unit 76, first multiplexors 120, and second multiplexors 122 is shown in accordance with one embodiment of microprocessor 10. A set of clock cycles are depicted, separated by dashed horizontal lines. The clock cycles are numbered 1, 2, and 3. For each clock cycle, the selections of first multiplexors 120 are shown via a block 140A-140D, and the selections of second multiplexors 122 are shown via a block 142A-142C. Instructions within a block are listed from right to left in program order (i.e. instruction F₁ in block 140A is prior to the other instructions in program order). Instructions are represented by an "F" for fast path instructions or an "M" for MROM instructions. A subscript is used to identify different fast path and MROM instructions.

During clock cycle 1, a set of potentially dispatchable instructions is selected via first multiplexors 120 (block 140A). It is noted that instructions F₂ and F₃ are shown in block 140A to depict instructions subsequent to instruction M₁. However, selection control unit 76 may not actually select instructions F₂ and F₃ since F₁ and M₁ comprise an allowable combination as shown in table 134. Alternatively, instructions F₂ and F₃ may be selected by first multiplexors 120C and 120D, but may not be selected by second multiplexors 122. Still further, the instructions may be routed through first multiplexors 120 and second multiplexors 122, but may be indicated to be invalid such that subsequent stages of the instruction processing pipeline ignore the instructions. As shown in block 142A, instructions F₁ and M₁ are selected by second multiplexors 122A and 122B, respectively. Second multiplexor 122C does not select a valid instruction since instructions F₁ and M₁ are concurrently dispatched. Additionally, selection control unit 76 sets the packed state, as shown in clock cycle 2. Since instructions F₁ and M₁ are speculatively dispatched concurrently, instruction M₁ is retained within byte queue 74 for potential redispach.

During clock cycle 2, instructions M₁, F₂, F₃, and F₄ are selected as a set of potentially dispatchable instructions (block 140B). Because the packed state is set, the double dispatch signal from MROM unit 34 is used to select which instructions from block 140B are dispatched. When the double dispatch signal is received, one of two possible sets of instructions are selected. If the double dispatch signal is deasserted, the instructions are selected as shown in block 142B. For this case, instruction M₁ was determined to not be a double dispatch instruction. Therefore, instruction M₁ is redispached. Additionally, other instructions are not concurrently dispatched with instruction M₁. Alternatively, the double dispatch signal may be asserted, resulting in the instruction selection shown in block 142C. Instruction M₁ is not redispached. Instead, instructions subsequent to M₁ are dispatched (i.e. instructions F₂, F₃, and F₄). Blocks 140C and 140D depict instructions selected as potentially dispatchable instructions during clock cycle 3 for the cases represented by blocks 142B and 142C, respectively.

Turning now to FIG. 11, a second example of operation of selection control unit 76, first multiplexors 120, and second multiplexors 122 is shown in accordance with one embodiment of microprocessor 10. For each clock cycle, the selections of first multiplexors 120 are shown via a block 144A-144D, and the selections of second multiplexors 122 are shown via a block 146A-146C. Instructions within a block are listed from right to left in program order (i.e. instruction M₁ in block 140A is prior to the other instructions in program order).

tions in program order). Instructions are represented by an "F" for fast path instructions or an "M" for MROM instructions. A subscript is used to identify different fast path and MROM instructions.

Clock cycle 1 in FIG. 11 is similar to clock cycle 1 in FIG. 10, except that instructions M_1 and F_1 are in reverse order for this example. Therefore, instruction F_1 is retained in byte queue 74 while being dispatched during clock cycle 1. Instruction M_1 is discarded, and the packed state is set for clock cycle 2. During clock cycle 2, instructions are selected via first multiplexors 120 as shown in block 144B. Since the packed state is set, one of two possible sets of instructions may be selected from the instructions in block 144B. Block 146B shows the instructions selected if the double dispatch signal is deasserted, while block 146C shows the instructions selected if the double dispatch signal is asserted. Blocks 144C and 144D depict instructions selected by first multiplexors 120 during clock cycle 3 for the cases shown in blocks 146B and 146C, respectively.

As block 146B of example of FIG. 11 shows, when the fast path instruction is the instruction redispached due to an unsuccessful concurrent dispatch of an MROM and fast path instruction, additional instructions may be dispatched as well. Advantageously, dispatch bandwidth may be maximized even during clock cycles in which a redispach is performed.

Turning now to FIG. 12, a computer system 200 including microprocessor 10 is shown. Computer system 200 further includes a bus bridge 202, a main memory 204, and a plurality of input/output (I/O) devices 206A-206N. Plurality of I/O devices 206A-206N will be collectively referred to as I/O devices 206. Microprocessor 10, bus bridge 202, and main memory 204 are coupled to a system bus 208. I/O devices 206 are coupled to an I/O bus 210 for communication with bus bridge 202.

Bus bridge 202 is provided to assist in communications between I/O devices 206 and devices coupled to system bus 208. I/O devices 206 typically require longer bus clock cycles than microprocessor 10 and other devices coupled to system bus 208. Therefore, bus bridge 202 provides a buffer between system bus 208 and input/output bus 210. Additionally, bus bridge 202 translates transactions from one bus protocol to another. In one embodiment, input/output bus 210 is an Enhanced Industry Standard Architecture (EISA) bus and bus bridge 202 translates from the system bus protocol to the EISA bus protocol. In another embodiment, input/output bus 210 is a Peripheral Component Interconnect (PCI) bus and bus bridge 202 translates from the system bus protocol to the PCI bus protocol. It is noted that many variations of system bus protocols exist. Microprocessor 10 may employ any suitable system bus protocol.

I/O devices 206 provide an interface between computer system 200 and other devices external to the computer system. Exemplary I/O devices include a modem, a serial or parallel port, a sound card, etc. I/O devices 206 may also be referred to as peripheral devices. Main memory 204 stores data and instructions for use by microprocessor 10. In one embodiment, main memory 204 includes at least one Dynamic Random Access Memory (DRAM) and a DRAM memory controller.

It is noted that although computer system 200 as shown in FIG. 12 includes one bus bridge 202, other embodiments of computer system 200 may include multiple bus bridges 202 for translating to multiple dissimilar or similar I/O bus protocols. Still further, a cache memory for enhancing the performance of computer system 200 by storing instructions

and data referenced by microprocessor 10 in a faster memory storage may be included. The cache memory may be inserted between microprocessor 10 and system bus 208, or may reside on system bus 208 in a "lookaside" configuration.

It is noted that, although double dispatch MROM instructions are described above with respect to dispatching microcode instructions concurrently with directly-decoded instructions, the present discussion applies to many different configurations. For example, for a microprocessor having four issue positions, microcode instructions which are parsed into two or three simpler instructions may be concurrently dispatched with two or one directly-decoded instructions, respectively. Such an embodiment might select a microcode instruction and two directly-decoded instructions for concurrent dispatch. If the microcode instruction is parsed into two instructions, then the concurrent dispatch is successful. If the microcode instruction is parsed into three instructions, then the concurrent dispatch of one of the directly-decoded instructions may be successful. Redispach of one of the concurrently dispatched instructions occurs. If the microcode instruction is parsed into four or more instructions, then the concurrent dispatch is unsuccessful and redispach of two of the concurrently dispatched instructions occurs. Similarly, additional issue positions may be added with extensions to the number of possible concurrent dispatches and the number of redispach scenarios. Any number of issue positions may be employed within various embodiments. Still further, although microcode instructions are divided depending upon the use of two issue positions or more than two issue positions, any division may be used. Continuing the four issue position example, microcode instructions may be specified as three dispatch or more than three dispatch. Microcode instructions which actually use two issue positions would waste an issue position, but the number of redispach scenarios is decreased. Still further, if a microcode instruction occupies each of the issue positions for several clock cycles, but the last clock cycle of instruction issue does not occupy each of the available issue positions, directly-decoded instructions subsequent to the microcode instruction may be dispatched during the last clock cycle of instruction issue by MROM unit 34 in response to the microcode instruction. In this case, a number of issue positions occupied is passed to selection control unit 76. Selection control unit 76 determines the number of instructions to be redispached from the number of instructions selected for dispatch and the number of issue positions occupied by the instructions issue by MROM unit 34.

Although the x86 microprocessor architecture and instruction set have been used as a specific example herein, it is noted that the apparatus and method described herein may be applicable to any microprocessor which employs microcode and directly-decoded instructions. Such embodiments are contemplated.

It is still further noted that the present discussion may refer to the assertion of various signals. As used herein, a signal is "asserted" if it conveys a value indicative of a particular condition. Conversely, a signal is "deasserted" if it conveys a value indicative of a lack of a particular condition. A signal may be defined to be asserted when it conveys a logical zero value or, conversely, when it conveys a logical one value. Additionally, various values have been described as being discarded in the above discussion. A value may be discarded in a number of manners, but generally involves modifying the value such that it is ignored by logic circuitry which receives the value. For example, if the value comprises a bit, the logic state of the

value may be inverted to discard the value. If the value is an n-bit value, one of the n-bit encodings may indicate that the value is invalid. Setting the value to the invalid encoding causes the value to be discarded. Additionally, an n-bit value may include a valid bit indicative, when set, that the n-bit value is valid. Resetting the valid bit may comprise discarding the value. Other methods of discarding a value may be used as well.

Table 1 below indicates fast path, double dispatch, and MROM instructions for one embodiment of microprocessor 10 employing the x86 instruction set:

TABLE 1

x86 Fast Path, Double Dispatch, and MROM Instructions	
X86 Instruction	Instruction Category
AAA	MROM
AAD	MROM
AAM	MROM
AAS	MROM
ADC	fast path
ADD	fast path
AND	fast path
ARPL	MROM
BOUND	MROM
BSF	fast path
BSR	fast path
BSWAP	MROM
BT	fast path
BTC	fast path
BTR	fast path
BTS	fast path
CALL	fast path/double dispatch
CBW	fast path
CWDE	fast path
CLC	fast path
CLD	fast path
CLI	MROM
CLTS	MROM
CMC	fast path
CMP	fast path
CMPS	MROM
CMPSB	MROM
CMPSW	MROM
CMPSD	MROM
CMPSQ	MROM
CMPSQ8B	MROM
CPUID	MROM
CWD	MROM
CWQ	MROM
DDA	MROM
DAS	MROM
DEC	fast path
DIV	MROM
ENTER	MROM
HLT	MROM
IDIV	MROM
IMUL	double dispatch
IN	MROM
INC	fast path
INS	MROM
INSB	MROM
INSW	MROM
INSD	MROM
INT	MROM
INTO	MROM
INVD	MROM
INVLPG	MROM
IRET	MROM
IRETD	MROM
Jcc	fast path
JCXZ	double dispatch
JECXZ	double dispatch
JMP	fast path
LAHF	fast path
LAR	MROM

TABLE 1-continued

x86 Fast Path, Double Dispatch, and MROM Instructions	
X86 Instruction	Instruction Category
LDS	MROM
LES	MROM
LFS	MROM
LGS	MROM
LSS	MROM
LEA	fast path
LEAVE	double dispatch
LGDT	MROM
LIDT	MROM
LLDT	MROM
LMSW	MROM
LODS	MROM
LODSB	MROM
LODSW	MROM
LODSD	MROM
LOOP	double dispatch
LOOPCond	MROM
LSL	MROM
LTR	MROM
MOV	fast path
MOVCC	fast path
MOVCR	MROM
MOVDR	MROM
MOVS	MROM
MOVSB	MROM
MOVSW	MROM
MOVSD	MROM
MOVSI	fast path
MOVZX	fast path
MUL	double dispatch
NEG	fast path
NOP	fast path
NOT	fast path
OR	fast path
OUT	MROM
OUTS	MROM
OUTSB	MROM
OUTSW	MROM
OUTSD	MROM
POP	double dispatch
POPA	MROM
POPAD	MROM
POPF	MROM
POPFD	MROM
PUSH	fast path/double dispatch
PUSHA	MROM
PUSHAD	MROM
PUSHF	fast path
PUSHFD	fast path
RCL	MROM
RCR	MROM
ROL	fast path
ROR	fast path
RDMSR	MROM
REP	MROM
REPE	MROM
REPZ	MROM
REPNE	MROM
REPNZ	MROM
RET	double dispatch
RSM	MROM
SAHF	fast path
SAL	fast path
SAR	fast path
SHL	fast path
SHR	fast path
SBB	fast path
SCAS	double dispatch
SCASB	MROM
SCASW	MROM
SCASD	MROM
SETcc	fast path
SGDT	MROM
SIDT	MROM

TABLE 1-continued

x86 Fast Path, Double Dispatch, and MROM Instructions	
X86 Instruction	Instruction Category
SHLD	MROM
SHRD	MROM
SLDT	MROM
SMSW	MROM
STC	fast path
STD	fast path
STI	MROM
STOS	MROM
STOSB	MROM
STOSW	MROM
STOSD	MROM
STR	MROM
SUB	fast path
TEST	fast path
VERR	MROM
VERW	MROM
WBINVD	MROM
WRMSR	MROM
XADD	MROM
XCHG	MROM
XLAT	fast path
XLATB	fast path
XOR	fast path

Note: Instructions including an SIB byte are also considered double dispatch instructions.

It is noted that a superscalar microprocessor in accordance with the foregoing may further employ the latching structures as disclosed within the co-pending, commonly assigned patent application entitled "Conditional Latching Mechanism and Pipelined Microprocessor Employing the Same", Ser. No. 08/400,608 filed Mar. 8, 1995, by Pflum et al. The disclosure of this patent application is incorporated herein by reference in its entirety.

It is further noted that aspects regarding array circuitry may be found in the co-pending, commonly assigned patent application entitled "High Performance Ram Array Circuit Employing Self-Time Clock Generator for Enabling Array Access", Ser. No. 08/473,103 filed Jun. 7, 1995 by Tran. The disclosure of this patent application is incorporated herein by reference in its entirety.

It is additionally noted that other aspects regarding superscalar microprocessors may be found in the following co-pending, commonly assigned patent applications: "Linearly Addressable Microprocessor Cache", Ser. No. 08/146,381, filed Oct. 29, 1993 by Witt; "Superscalar Microprocessor Including a High Performance Instruction Alignment Unit", Ser. No. 08/377,843, filed Jan. 25, 1995 by Witt, et al; "A Way Prediction Structure", Ser. No. 08/522,181, filed Aug. 31, 1995 by Roberts, et al; "A Data Cache Capable of Performing Store Accesses in a Single Clock Cycle", Ser. No. 08/521,627, filed Aug. 31, 1995 by Witt, et al; "A Parallel and Scalable Instruction Scanning Unit", Ser. No. 08/475,400, filed Jun. 7, 1995 by Narayan; and "An Apparatus and Method for Aligning Variable-Byte Length Instructions to a Plurality of Issue Positions", Ser. No. 08/582,473, filed Jan. 2, 1996 by Narayan, et al. The disclosure of these patent applications are incorporated herein by reference in their entirety.

In accordance with the above disclosure, a method and apparatus for concurrently dispatching microcode instructions and directly-decoded instructions is provided. The microcode and directly-decoded instructions are speculatively dispatched under the assumption that the microcode instruction occupies a fixed, predetermined number of issue

positions. The predetermined number of issue positions is less than the total number of issue positions available within the microprocessor. If the microcode instruction is found to occupy a larger number of issue positions, then one or more of the concurrently dispatched instructions are redispached in the subsequent clock cycle. Advantageously, dispatch bandwidth is increased during clock cycles in which a microcode instruction occupying the predetermined fixed number of issue positions is concurrently dispatched along with additional instructions.

Numerous variations and modifications will become apparent to those skilled in the art once the above disclosure is fully appreciated. It is intended that the following claims be interpreted to embrace all such variations and modifications.

What is claimed is:

1. A microprocessor comprising:

an instruction cache configured to store instructions;

an instruction alignment unit coupled to receive a plurality of instructions fetched from said instruction cache,

said plurality of instructions including a directly-decoded instruction and a microcode instruction, wherein said instruction alignment unit is configured to select a first dispatch plurality of instructions including said directly-decoded instruction and said microcode instruction, and wherein said instruction alignment unit is configured to select said first dispatch plurality of instructions from said plurality of instructions; and

a microcode unit coupled to receive said microcode instruction from said instruction cache, wherein said microcode unit is configured to determine a number of directly-decoded instructions corresponding to said microcode instruction, and wherein said microcode unit is configured to transmit a signal indicative of said number of directly-decoded instructions;

wherein said instruction alignment unit is coupled to receive said signal from said microcode unit, and wherein said instruction alignment unit is configured to determine if said first dispatch plurality of instructions is concurrently dispatchable in response to said signal, and wherein said instruction alignment unit is configured to discard one of said microcode instruction and said directly-decoded instruction from said first dispatch plurality of instructions in response to determining that said first dispatch plurality of instructions is not concurrently dispatchable.

2. The microprocessor as recited in claim 1 further comprising a plurality of decode units coupled to receive said first dispatch plurality of instructions from said instruction alignment unit, wherein said plurality of decode units are configured to decode directly-decoded instructions.

3. The microprocessor as recited in claim 1 wherein said instruction alignment unit is further configured to retain said one of said microcode instruction and said directly-decoded instruction discarded from said first dispatch plurality of instructions, and wherein said instruction alignment unit is further configured to subsequently dispatch said one of said microcode instruction and said directly-decoded instruction.

4. The microprocessor as recited in claim 3 wherein said one of said microcode instruction and said directly-decoded instruction is a second one in program order of said microcode instruction and said directly-decoded instruction.

5. The microprocessor as recited in claim 4 wherein said instruction alignment unit is configured to select a second dispatch plurality of instructions if said second one in program order is said directly-decoded instruction, and wherein said directly-decoded instruction is included within said second dispatch plurality of instructions.

33

6. The microprocessor as recited in claim 4 wherein said instruction alignment unit is configured to dispatch said microcode instruction individually if said second one of said microcode instruction and said directly-decoded instruction is said microcode instruction.

7. The microprocessor as recited in claim 3 wherein said instruction alignment unit comprises an instruction queue configured to store said plurality of instructions fetched from said instruction cache, and wherein said instruction alignment unit is configured to retain said one of said microcode instruction and said directly-decoded instruction by inhibiting deletion of said one of said microcode instruction and said directly-decoded instruction from said instruction queue.

8. The microprocessor as recited in claim 1 wherein said instruction alignment unit comprises a first alignment stage and a second alignment stage.

9. The microprocessor as recited in claim 8 wherein said first alignment stage is configured to select said first dispatch plurality of instructions.

10. The microprocessor as recited in claim 9 wherein said second alignment stage is coupled to receive said first dispatch plurality of instructions from said first alignment stage, and wherein said second alignment stage is coupled to receive said signal from said microcode unit and to discard said one of said microcode instruction and said directly-decoded instruction from said first dispatch plurality of instructions.

11. The microprocessor as recited in claim 10 wherein said first alignment stage is coupled to receive said signal from said microcode unit and to redispach said one of said microcode instruction and said directly-decoded instruction, wherein said one of said microcode instruction and said directly-decoded instruction is a second one in program order of said microcode instruction and said directly-decoded instruction.

12. The microprocessor as recited in claim 1 wherein said instruction cache comprises an instruction scan unit configured to scan said plurality of instructions fetched from said instruction cache, and wherein said instruction scan unit is configured to detect said microcode instruction and to route said microcode instruction to said microcode unit.

13. A computer system comprising:

a microprocessor including:

an instruction cache configured to store instructions;
an instruction alignment unit coupled to receive a plurality of instructions fetched from said instruction cache, said plurality of instructions including a directly-decoded instruction and a microcode instruction, wherein said instruction alignment unit is configured to select a first dispatch plurality of instructions including said directly-decoded instruction and said microcode instruction, and wherein said instruction alignment unit is configured to select said first dispatch plurality of instructions from said plurality of instructions; and

a microcode unit coupled to receive said microcode instruction from said instruction cache, wherein said microcode unit is configured to determine a number of directly-decoded instructions corresponding to said microcode instruction, and wherein said micro-

34

code unit is configured to transmit a signal indicative of said number of directly-decoded instructions;

wherein said instruction alignment unit is coupled to receive said signal from said microcode unit, and wherein said instruction alignment unit is configured to determine if said first dispatch plurality of instructions is concurrently dispatchable in response to said signal, and wherein said instruction alignment unit is configured to discard one of said microcode instruction and said directly-decoded instruction from said first dispatch plurality of instructions in response to determining that said first dispatch plurality of instructions is not concurrently dispatchable; and

an input/output (I/O) device coupled to said microprocessor and to another computer system, wherein said I/O device is configured to communicate between said computer system and said another computer system.

14. The computer system as recited in claim 13 wherein said I/O device comprises a modem.

15. A method for dispatching instructions in a microprocessor, the method comprising:

speculatively packing a microcode instruction and a directly-decoded instruction into a first dispatch plurality of instructions for dispatch to a plurality of decode units;

determining a number of directly-decoded instructions corresponding to said microcode instruction;

determining if said first dispatch plurality of instructions are concurrently dispatchable responsive to said determining a number of directly-decoded instructions corresponding to said microcode instruction;

discarding one of said microcode instruction and said directly-decoded instruction from said first dispatch plurality of instructions responsive to said determining if said first dispatch plurality of instructions are concurrently dispatchable; and

dispatching said first dispatch plurality of instructions subsequent to said discarding.

16. The method as recited in claim 15 further comprising redispaching said one of said microcode instruction and said directly-decode instruction responsive to said discarding.

17. The method as recited in claim 15 wherein said discarding comprises discarding said microcode instruction if said microcode instruction is subsequent to said directly-decoded instruction in program order.

18. The method as recited in claim 15 wherein said discarding comprises discarding said directly-decoded instruction if said directly-decoded instruction is subsequent to said microcode instruction in program order.

19. The method as recited in claim 15 wherein said determining if said first dispatch plurality of instructions are concurrently dispatchable comprises determining if a sum of said number of said directly-decoded instructions corresponding to said microcode instruction and a number of remaining ones of said first dispatch plurality of instructions is less than a number of said plurality of decode units.

* * * * *